NEGATIVE BINOMIAL-GENERALIZED EXPONENTIAL DISTRIBUTION: GENERALIZED LINEAR MODEL AND ITS APPLICATIONS

A Thesis

by

PRATHYUSHA VANGALA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Dominique Lord
Co-Chair of Committee,	Yunlong Zhang
Committee Member,	Daren B.H Cline
Head of Department,	Robin Autenrieth

May 2015

Major Subject: Civil Engineering

ABSTRACT

Modelling crash data has been an integral part of the research done in highway safety. Different tools have been suggested by researchers to analyze crash data. One such tool, which was recently proposed, is the Negative Binomial Generalized Exponential (NB-GE) distribution. As the name suggests, it is a combination of Negative Binomial and Generalized Exponential distribution. This distribution has three parameters and can handle over-dispersed crash data which are characterized by a large number of zeros and/or long tail. This research seeks to develop a generalized linear model (GLM) for NB-GE distribution and discuss its applications in crash data analysis. The NB-GE GLM was applied to two over-dispersed crash datasets and its performance was compared to Negative Binomial-Lindley (NB-L) and Negative Binomial (NB) models using various statistical measures. It was found that NB-GE performs almost as well as NB-L model and performs much better than the NB model. This research tried to determine the percentage of zeroes and the dispersion in the dataset where the NB-GE model is recommended over the NB model for ranking sites. Datasets were simulated for different scenarios. It was found that for high dispersion the NB-GE model performs better than the NB model when the percentage of zero counts in the dataset is greater than 80%. When dataset has lower than 80% zeroes then NB model and NB-GE model perform similarly. Hence for lower percentages NB model would be preferred as it is simpler and easier to use.

DEDICATION

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank everyone who helped me with my research.

I would like to thank my advisor, Dr. Dominique Lord, with a deep sense of gratitude for noticing my interest in statistics and assigning me a challenging project and guiding me. I would also like to thank him for helping me to get in touch with experts in various fields when needed, and providing me with the resources to help with my simulations when I needed the most.

I would like to thank my mentor, Dr. Srinivas Geedipally (Assistant Research Engineer at TTI), who constantly guided me, gave valuable suggestions through-out the project. I would also like to thank him for taking time from his busy schedule to discuss with me on a regular basis and even over the weekends.

I would also like to thank Dr. Soma S. Dhavala for helping me with coding NB-GE GLM in OpenBUGS.

I would like to thank Dr. Yunlong Zhang and Dr. Daren B.H Cline for agreeing to be on my MS thesis committee.

I would also like to thank my family for the confidence they had in me without which, I would have never been able to pursue and complete my master's study.

I would like to thank Texas A&M University and the Department of Civil Engineering for giving me an opportunity to pursue my master's degree.

I am grateful to God without whose blessings I would not have been here.

NOMENCLATURE

AADT	Annual Average Daily Traffic
BUGS	Bayesian inference Using Gibbs Sampling
COM-POISSON	Conway-Maxwell-Poisson
CURE	Cumulative Residual
GLM	Generalized Linear Model
MAD	Mean Absolute Deviance
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood estimation
MSPE	Mean Squared Predictive Error
NB	Negative Binomial
NB-CR	Negative Binomial-Crack
NB-GE	Negative Binomial-Generalized Exponential
NB-L	Negative Binomial-Lindley
PDF	Probability Density Function
PIG	Poisson Inverse Gaussian
PMF	Probability Mass Function
SI	Sichel

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

NOMENCLATURE v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES ix

1. INTRODUCTION 1

    1.1 Study Objectives 4
    1.2 Outline of the Thesis 5

2. BACKGROUND 7

    2.1 Poisson Model 7
    2.2 Negative Binomial Model 9
    2.3 Previous Studies 10
    2.4 Zero-Inflated Model 12
    2.5 Negative Binomial-Lindley Model 13
    2.6 Poisson Inverse Gaussian (PIG) Distribution 14
    2.7 Sichel (SI) Distribution 16
    2.8 NB-Crack Distribution 18
    2.9 Other Models 19
        2.9.1 Gamma Count Model 20
        2.9.2 Conway-Maxwell-Poisson Model 21
    2.10 Negative Binomial-Generalized Exponential Model 23
    2.11 Estimation Methods 25
    2.12 Summary 28

3. PERFORMANCE OF THE NB-GE DISTRIBUTION 29

3.1 Description of Datasets 29
3.2 Goodness of Fit 31
3.3 Results and Discussion 32
3.4 Summary 37

4. APPLICATION OF THE NB-GE GENERALIZED LINEAR MODEL FOR OVER-DISPERSED CRASH DATA 38

4.1 NB-GE Generalized Linear Model 38
4.2 Description of Datasets 39
4.3 Goodness of Fit 41
4.4 Results and Discussion 42
4.5 Summary 49

5. PERFORMANCE OF NB-GE GENERALIZED LINEAR MODEL FOR SIMULATED CRASH DATA 51

5.1 Data Simulation 51
5.2 Performance Measures 52
5.3 Results and Discussion 55
5.4 Summary 62

6. SUMMARY AND RECOMMENDATIONS 64

6.1 Summary 64
6.2 Recommendations 65

REFERENCES 65

APPENDIX A 76

APPENDIX B 79

LIST OF FIGURES

Page

Figure 1. Predicted and observed zero counts for the first dataset 34

Figure 2. Predicted and observed zero counts for the second dataset 36

Figure 3. Cumulative residual plot for Indiana data (ADT variable) 44

Figure 4. Cumulative residual plot for Indiana data (Friction variable) 45

Figure 5. Cumulative residual plot for Michigan data (AADT variable) 47

Figure 6. Cumulative residual plot for Michigan data (length segment) 48

LIST OF TABLES

Page

Table 1. Summary statistics for single-vehicle fatal crashes on divided multilane rural highways between 1997 and 2001 30

Table 2. Summary statistics for single-vehicle roadway departure crashes on rural two-lane horizontal curves between 2003 and 2008..... 30

Table 3. Single-vehicle fatal crashes on divided multilane rural highways between 1997 and 2001 33

Table 4. Single-vehicle roadway departure crashes on rural two-lane horizontal curves between 2003 and 2008 35

Table 5. Summary statistics for Indiana data 40

Table 6. Summary statistics for Michigan data.....41

Table 7. Modelling results for Indiana data 43

Table 8. Modelling results for Michigan data 46

Table 9. Possible outcomes of classification (Miranda-Moreno, 2006).....53

Table 10. Number of mis-specified sites for different scenarios.....55

Table 11. Performance measures for 50% zero counts and low dispersion.....56

Table 12. Performance measures for 50% zero counts and high dispersion.....57

Table 13. Performance measures for 60% zero counts and low dispersion.....58

Table 14. Performance measures for 60% zero counts and high dispersion.....58

Table 15. Performance measures for 70% zero counts and low dispersion.....59

Table 16. Performance measures for 70% zero counts and high dispersion.....59

Table 17. Performance measures for 80% zero counts and low dispersion.....60

Table 18. Performance measures for 80% zero counts and high dispersion.....61

Table 19. Performance measures for 90% zero counts and low dispersion.....61

Table 20. Performance measures for 90% zero counts and high dispersion.....62

1. INTRODUCTION

Many researchers have sought to examine relationships between explanatory variables, such as traffic and roadway characteristics, and crashes. Modelling crash datasets based on their characteristics is the cornerstone of highway safety. There are different factors that might contribute to or influence motor vehicle crashes. These factors need to be taken into account while modeling this kind of dataset. Some analysts have used only traffic flow data to analyze crashes (Maher, 1991; Shankar et al., 1995; Golob and Recker, 1987; Miaou, 1994) while others (Abdel-Aty and Abdalla, 2004) have used more extensive databases, which included both geometric characteristics of roadway and real time traffic characteristics to predict crashes among others. Lord and Mannering (2010) have provided a summary of the latest models used for this purpose. Additional discussion about these latest models can be found in Mannering and Bhat (2014).

High degree of randomness can be found in crash data. Therefore, based on the characteristics of the data, a suitable model needs to be chosen for making proper inferences. The Poisson model can usually handle count data whose mean is equal to the variance. However, it has been shown that crash data are usually characterized by over-dispersion, which means that the sample variance is larger than the sample mean. When this happens, the Poisson model cannot handle such datasets as it may consider certain covariates to have significant influence though they do not affect the data (Park and Lord, 2007; Hilbe, 2011). To overcome this problem, the negative binomial (NB) model

has been proposed and is now considered the widely accepted model for crash data analysis (Lord and Mannering, 2010). Recently, new and innovative distributions and models have been proposed to handle over-dispersed and, sometimes, under-dispersed data. They include the Conway-Maxwell-Poisson (COM-Poisson) (Conway and Maxwell, 1962; Shmueli, 2005), Double-Poisson (DP) (Efron, 1986; Zou et al., 2014) and Gamma models (Oh et al., 2006).

Many datasets in highway safety have been characterized by a large number of zeroes or highly dispersed data. According to Lord et al., (2005 & 2007) such characteristics might be observed due to the following factors:

1. The sites which have a combination of high heterogeneity, low exposure and sites categorized as high risk.
2. The analysis is conducted with short time or small spatial scales.
3. A considerable fraction of data contains missing or mis-reported crashes.
4. Critical variables were not considered in the crash prediction models.

Over the years, a few models such as the Zero Inflated (Shankar et al., 1997; Shankar et al., 2001) and the Negative Binomial- Lindley (Geedipally et al., 2012; Hallmark et al., 2013) have been suggested by researchers for over-dispersed crash data. Zero Inflated models are based on the assumption that crash datasets can be divided into two states: safe and non-safe states. Although they have been reported to fit well, Lord et al., (2005 & 2007) have argued that considering a dual state process may not be the correct way to approach large number of zero counts in crash data.

To overcome this problem, some researchers in highway safety have been examining the application of three-parameter models to evaluate datasets with excess zeroes. Lord and Geedipally (2011), for example, analyzed crash data containing large number of zeroes using the Negative Binomial-Lindley (NB-L) distribution and compared it with the performance of Negative Binomial. The NB-L is a distribution that combines or mixes the Negative binomial and Lindley distributions. The authors noted that the NB-L performs much better than NB for data characterized by large number of zeroes. The same characteristic in terms of performance was noted for the NB-L generalized linear model (GLM) (Geedipally et al., 2012). The NB-GE distribution (Aryuyuen and Bodhisuwan, 2013) is one such distribution that was also recently developed to handle over-dispersed crash datasets with large number of zeroes. The emphasis of this proposed thesis is to develop the NB-GE GLM and compare its performance to the NB-L and NB models.

Over the past few years, Bayes methods are being preferred over traditional Maximum Likelihood Estimation (MLE) to estimate coefficients for models, particularly likelihood is very complex to analyze. There are some open source software programs available, such as WinBUGS and OpenBUGS (Spiegelhalter et al., 2003) that use Bayesian methods for estimating coefficients of regression models. BUGS stands for Bayesian inference Using Gibbs Sampling. These software programs use Markov Chain Monte Carlo (MCMC) techniques to estimate the coefficients of statistical models (Lunn et al., 2000). OpenBUGS will be used to estimate the coefficients for NB-GE GLM in this research because it has a built-in function for the NB and GE distributions

1.1 Study Objectives

Crash data are usually over-dispersed and most datasets are characterized with many zeroes. The NB model is currently the most widely used for analyzing dispersed datasets, but is found to be inadequate when the data contain excess zeros and/or is highly over-dispersed. The focus of this thesis is to examine the NB-GE model, which was documented to perform well when the data contains a large amount of zeros, and in the process achieve the following objectives:

1. Examine the performance of the NB-GE distribution for over-dispersed crash datasets with large number of zeroes and compare it with NB and NB-L distributions. The performance of NB-GE distribution will be examined and compared to other distributions such as the Poisson, NB and NB-L. Different goodness-of-fit (GOF) statistics will be used as performance measures. Two datasets that have been collected as a part of a previous National Cooperative Highway Research Program (NCHRP) project will be used. These datasets are chosen as they are characterized by over-dispersion and excess zeroes.

2. Develop a NB-GE GLM for analyzing over dispersed crash datasets using the Bayesian approach. A GLM for NB-GE will be developed which will later be used for its performance analysis.

3. Apply the GLM to crash datasets and compare its performance to NB and NB-L models. Two over-dispersed crash datasets will be used for this purpose and its performance will be compared using different GOF statistics.

4. Examine the properties of NB-GE GLM for different percentages of zeroes and different dispersion levels in the data. In order to accomplish this, data with varying percentage of zero counts and different dispersion levels is first simulated. The simulation protocol followed is described later in the documentation. The performance of NB-GE GLM for different scenarios is examined and compared to NB model. Ranking the sites for hot spot identification is considered as the performance measure.

1.2 Outline of the Thesis

This subsection provides a brief outline of the Thesis.

Section 2 of this thesis reports different models that have been used to analyze crash data. A brief overview of different distributions such as the Poisson, negative binomial, Gamma, zero-Inflated, COM-Poisson, Poisson inverse Gaussian, Sichel, negative binomial- Lindley and negative binomial-generalized exponential distribution are provided. Applications and limitations of these models are also discussed. A summary of the estimation methods is also provided.

Section 3 documents the performance of the NB-GE distribution using two different datasets which were collected as a part of NCHRP 17-29 research project titled "*Methodology for estimating the safety performance of multilane rural highways*" (Lord et al., 2008). Its performance is compared to the Poisson, NB and NB-GE distributions using different GOF tests such as chi-squared and log-likelihood.

Section 4 presents the performance analysis of the NB-GE GLM using two datasets (Geedipally et al., 2012). Mean absolute deviation (MAD), Mean square

predicted error (MSPE) and Cumulative residual (CURE) plots were used (Oh et al., 2003; Lord et al., 2007). The performance of NB-GE model is compared to NB-L and NB models.

Section 5 summarizes the performance of NB-GE GLM using simulated data. Data with different percentage of zero counts and different dispersion values are simulated. Using these data, the performance of both NB and NB-GE models is compared. Ranking the sites for hot spot identification is considered as a performance measure.

Section 6 documents the findings of this research and also includes recommendations for future work.

2. BACKGROUND

Different distributions have been proposed by many researchers for analyzing crash data. A brief overview of different statistical distributions and models that were proposed to handle over-dispersed and under-dispersed crash data are provided in this section. The limitations of the models are also discussed. At the end of this section, a brief introduction to the NB-GE distribution is provided.

2.1 Poisson Model

The Poisson distribution is used to find the probability of occurrence of a certain number of events in a fixed amount of time or space. Hence, it is a discrete probability distribution. The occurrences of events are considered to be independent of each other. Rareness, discreteness and randomness are the main characteristics of crashes. According to Lord et al. (2005), crash data can be best characterized as Bernoulli trails which have low probability and large number. Such probability model that accounts for a series of Bernoulli trials is known as the binomial distribution. It is given by:

$$P(Z = n) = \binom{N}{n} p^n (1-p)^{N-n} \quad \text{.....(1)}$$

Where n = 0, 1 ,2, ..., N.

Here, n is the number of crashes.

Its mean is given by $E(Z) = Np$.....(2)

Variance is given by $V(Z) = Np(1-p)$.....(3)

Due to the low probability and large number, the number of crashes can be characterized as Poisson trials and can be approximated by the Poisson distribution.

The Probability mass function (PMF) of Poisson distribution is given by the following equation (4):

$P(y_i/\lambda_i) = \exp(-\lambda_i)\, \lambda_i y_i\ /y_i!$.....(4)

Where,

y_i is the number of crashes at site i;

λ_i is the mean of crashes at site i.

The mean and variance of the Poisson distribution is given by:

$E(y_i) = \lambda_i$.....(5)

$Var(y_i) = \lambda_i$.....(6)

One of the limitations of the Poisson distribution and model is that the variance should be equal to its mean. Crash data are often characterized by over-dispersion (variance greater than mean) and under-dispersion (variance less than mean) in some cases. Over dispersion of crash data is observed as variables (Lord and Park, 2008) can be characterized by uncertainties and unobserved differences that exist among sites (Washington et al., 2003). On rare occasions, crash data have shown under-dispersion and often low sample mean was found to be the cause (Lord and Mannering, 2010). Inconsistency in standard error of the parameter estimates will be observed if the Poisson distribution is used to analyze over-dispersed and under-dispersed crash data (Cameron and Trivedi, 1998). Using the Poisson distribution for such crash data might also result in the under-estimation of standard errors (Miranda-Moreno, 2006; Park and Lord, 2007).

2.2 Negative Binomial Model

The Negative binomial (Poisson-gamma) distribution is mostly used to analyze crash data which exhibit over-dispersion (Lord and Mannering, 2010). The mathematics involved in finding the relationship between mean and variance was also found to be simple (Hauer, 1997). Various statistical software programs such as R (Venables et al., 2005), WinBUGS (Spiegelhalter et al., 2003) and OpenBUGS already have built-in functions to estimate NB models.

The NB distribution is a discrete probability distribution of successes in a sequence of Bernoulli trials before a predetermined number of failures occur.

The PMF of NB distribution is given by:

$$P(Y=y;r, p)=((y+r-1)|y)(1-p)^r p^y \quad \text{.....(7)}$$

Where,

p = probability of success in each trial;

r = number of failures; and

y = number of success

The probability of success in each trial, p, is given by:

$$p = \frac{ø}{\mu+ø} \quad \text{.....(8)}$$

Where,

ø = inverse of the dispersion parameter (α);

μ = mean of the observations;

From the above equations, PMF of NB distribution can be re-written as:

$$P(Y=y,\mu,\phi)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} \quad \text{.....(9)}$$

In the NB regression model, the mean is related to the covariates using the following equation:

$$\mu = \exp(\sum x\beta) \quad \text{.....(10)}$$

One of the disadvantages of NB model is that it cannot handle under-dispersed data very well or not at all. Theoretically, NB model can handle under-dispersion if the dispersion parameter is negative. In this case, the conditioned mean of Poisson would not follow gamma distribution. The parameter estimates would not be reliable (Lord et al., 2010) and the PDF of the distribution would be mis-specified (Clark and Perry, 1989; Saha and Paul, 2005).

2.3 Previous Studies

Several studies have documented the use of different models to handle count data containing excess zeroes. Lambert (1992) has suggested the Zero inflated Poisson (ZIP) model for such data in manufacturing. According to this paper, it was assumed that a zero-defect equipment is manufactured with a probability p and an equipment with some defects is manufactured with a probability 1-p and the defects follow Poisson process with a mean λ. Here, p may or may not be a function of λ. Though its interpretation was easy it was found that ZIP regression inflates the zero counts.

ZIP model was first examined by Miaou (1994) for modelling crash data. Poisson, NB and ZIP models were evaluated to find the relationship between crash

counts and the design variables of road section. It was suggested that when the data exhibits over-dispersion both the NB and ZIP should be considered. When the data contains excess zeroes in addition to over-dispersion ZIP should be considered though it is difficult to interpret when compared to the NB.

Another study also looked into the applicability of ZIP and ZINB distributions to crash frequencies (Shankar et al., 1997). For minor-arterial data, ZINB distribution was found to be appropriate while for collector-arterial data, ZIP suited the best. NB distribution was found to be appropriate for principal-arterial data. It was also found that the ZIP process provides flexibility. It also provides the opportunity to find the design factors that contribute to the crash occurrence.

The applicability of the ZIP model was also tested for pedestrian-traffic crashes (Shankar et al., 2003). The ZIP model was found to be the most suited model for crashes involving pedestrians. The roadway characteristics that play significant role in such crashes were also detected.

The ZIP regression model was used to determine the factors that increase the probability of accident occurrence at signalized tee intersections having excess zeroes (Kumara and Chin, 2003). It was found that left-turn volumes, uncontrolled left-turn slip roads, signal phases per cycle, existence of horizontal curves, and permissive right-turn phases resulted in an increase in the number of crashes. These models are briefly discussed in the next few sections.

2.4 Zero-Inflated Model

The Poisson and NB model tend to produce biased estimates for over-dispersed crash data that contain excess zeroes. The Zero-Inflated or Zero-altered model was proposed by researchers to handle such over-dispersed crash data (Miaou, 1994; Shankar et al, 1997 and 2003; Qin et al, 2004). In this model, a dual-state process is considered in which the crash datasets are divided into two states: safe state (zero-count state) and non-safe state (non-zero state). The sites that have a very low or zero probability of accidents occurring were classified under safe state and the sites where the crashes follow a Poisson or NB distribution were classified under non-safe state. The corresponding Zero-inflated probability models are called the Zero-inflated Poisson (ZIP) model and the Zero-inflated negative binomial (ZINB) model.

The PDF for the ZIP model is given by:

$$P(Y)= \delta+(1-\delta)e^{-\lambda};Y=0 \quad (11)$$

$$P(Y)=(1-\delta)\frac{e^{-\lambda}\lambda^{y}}{y!};Y\geq 0 \quad (12)$$

Where,

Y = number of crashes on the road segment,

δ = probability of zero crash state on the road segment,

1-δ = probability of crashes following Poisson distribution.

The PDF for ZINB is given by:

$$P(Y)= \delta+(1-\delta)(1+\alpha\lambda)^{-\alpha^{-1}};Y=0 \quad (13)$$

$$P(Y)=(1-\delta)\left[\frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!}\ (\lambda\alpha)^{y}(1+\alpha\lambda)^{-(y+\alpha^{-1})}\right];Y\geq 0 \quad (14)$$

Where,

Y = number of crashes on the road segment,

δ = probability of zero crash state,

1-δ = probability of crashes following Negative binomial distribution,

α = dispersion parameter and

λ = mean.

One of the limitations of this model is that the safe state has a long term mean equal to zero (Lord et al., 2005b; Warton, 2005), which is theoretically not feasible with crash data. The risk of a crash cannot be zero, unless nobody uses the facility (i.e., intersection, roadway segment, etc.).

2.5 Negative Binomial-Lindley Model

The Negative binomial-Lindley (NB-L) distribution is a three-parameter distribution and is a combination of the NB and Lindley distributions. It was recently introduced to analyze crash data (Lord and Geedipally, 2011).

The PMF of the NB distribution is the same as before:

$$P(Y=y,\mu,\phi)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} \quad \text{.....(15)}$$

Where,

μ= mean response and,

ø= inverse of the dispersion parameter α.

The PMF of NB-L distribution can be written as:

$$P(Y=y,\ \mu,\ \text{ø},\ \theta)=\int NB(y;\ \text{ø},\ \varepsilon\mu)\,Lindley(\varepsilon;\ \theta)\,d\varepsilon \quad \text{.....(16)}$$

Here, ε follows Lindley distribution and y follows NB distribution.

Lindley distribution is a combination of exponential and gamma distributions (Lindley, 1958; Lord and Geedipally, 2011). Its PMF is given by:

$$f(X=x;\theta)=\frac{\theta^2}{1+\theta}(1+x)e^{-\theta x};\ \theta>0,\ x>0 \quad \text{.....(17)}$$

Geedipally et al. (2011) have evaluated the performance of NB-L model with respect to crash data that exhibit high dispersion with excess zeroes and compared its performance to the ZINB and NB models. It was found that the NB-L distribution and its GLM perform better and provide a good statistical fit for crash data which exhibit such characteristics when compared to NB model.

2.6 Poisson Inverse Gaussian (PIG) Distribution

This distribution is a combination of Poisson and Inverse Gaussian distribution (Zha et al, 2014). The flexibility of inverse Gaussian distribution helps in handling data which exhibit high dispersion. It is a special type of SI distribution which is obtained by setting a value of -0.5 for shape parameter and has only two parameters.

The crash mean on any segment i, Y_i is assumed to follow Poisson distribution (Miaou and Lord, 2003):

$$Y_i|\mu_i \sim \text{Poisson}(\mu_i) \qquad i=1,2,\ldots n \quad \text{.....(18)}$$

$$\mu_i=E(Y_i|\mu_i)=\text{Var}(Y_i|\mu_i)=f(X;\beta)=EXP(X\beta) \quad \text{.....(19)}$$

Where,

μ_i is the mean,

f(.) is the Link functional form,

X is the covariate vector,

β is the vector of regression parameters.

An error term is introduced to account for over-dispersion that is likely to be observed in the crash data as given below:

$$EXP(X^T\beta+\varepsilon_i)=\mu_i EXP(\varepsilon_i)=\mu_i \nu_I) \quad (20)$$

If $g(\nu_i)$ is the PDF of ν_i, the marginal distribution for Y_i is given by:

$$P(Y_i=y_i|\mu_i)=\int f(y_i|\mu_i, \nu_i)g(\nu_i)d\nu_i \quad (21)$$

It is assumed that ν_i follows Inverse Gaussian distribution and is independent of covariates. The mean and shape parameter are assumed to be equal to 1 and $1/\tau$. The PDF of ν_i is given by (Stasinopoulos and Rigby, 2007):

$$g(\nu_i)=(2\pi\tau\nu_i^3)^{-0.5}e^{-(\nu_i-1)^2/2\tau\nu_i}, \quad \nu_i>0 \quad (22)$$

Where,

$\tau = Var(\nu_i)$,

$E(\nu_i) = 1$.

The PIG distribution, $PIG(\mu_i, \tau)$, is given by:

$$P(y_i|\mu_i, \tau)=\left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}} \frac{\mu_i^{y_i} e^{\frac{1}{\tau}} K_{y_i-\frac{1}{2}}(\alpha)}{(\alpha_i\tau)^{y_i} y_i!} \quad (23)$$

Where,

$$\alpha_i=\sqrt{\frac{1}{\tau^2}+\frac{2\mu_i}{\tau}},$$

$$K_\lambda(t)=\frac{1}{2}\int_0^\infty x^{\lambda-1}e^{\left\{-\frac{1}{2}t\left(x+x^{-1}\right)\right\}}dx \quad (24)$$

The mean and variance of PIG distribution are given by:

$$E(Y_i)=E\{E(Y_i|\mu_i\nu_i)\}=E(\mu_i\nu_i)=\mu_i \quad (25)$$

$$Var(Y_i)=Var\{E(Y_i|\mu_i\nu_i)\}+E\{Var(Y_i|\mu_i\nu_i)\}=\mu_i+\tau\mu_i^2 \quad (26)$$

The PIG model has been used by researchers as an alternative to NB model for count data (Willmot, 1987, Shoukri et al., 2004, Dean et al., 1989, Jagger and Elsner, 2012). Zha et al. (2014) analyzed the PIG model for crash data and compared its performance to NB model by introducing varying dispersion parameter. Based on GOF statistics, it was found that the PIG model performs better than the NB model. The computational effort required for parameter estimation was also found to be very less when compared to the recently introduced NB-L model. Hence, the PIG model was suggested as an alternative for NB model for analyzing crash data.

2.7 Sichel (SI) Distribution

The SI distribution is a combination of Poisson distributions and is used to analyze data that exhibit high dispersion. The Poisson rate is assumed to have an inverse Gaussian distribution (GIG). It was used by some researchers (Zou et al., 2011, Zou et al., 2012 and Wu et al., 2013) to analyze crash data. The PDF of the GIG distribution is given by (Stasinopoulos and Rigby, 2007):

$$f(\lambda|\mu,\sigma,\nu)=\left(\frac{c}{\mu}\right)^{\nu}\left[\frac{\lambda^{\nu-1}}{2K_{\nu}\left(\frac{1}{\sigma}\right)}\right]\exp\left[-\frac{1}{2\sigma}\left(\frac{c\lambda}{\mu}+\frac{\mu}{c\lambda}\right)\right] \quad (27)$$

The number of crashes, y, is given by the following equation:

$$p(y|\mu,\sigma,\nu)=\int_0^{\infty} p(y|\lambda)f(\lambda|\mu,\sigma,\nu)d\lambda \quad (28)$$

The PDF obtained after solving the integral is given by (Wu et al., 2014):

$$p(y|\ \mu,\sigma,\nu)=\frac{\left(\frac{\mu}{c}\right)^{y} K_{y+\nu}(\alpha)}{K_{\nu}\left(\frac{1}{\sigma}\right) y!\ (\alpha\sigma)^{y+\nu}} \quad \text{.....(29)}$$

$$\alpha^{2}=\sigma^{-2}+2\mu(c\sigma)^{-1} \quad \text{.....(30)}$$

$$c=\frac{K_{\nu+1}(\frac{1}{\sigma})}{K_{\nu}(\frac{1}{\sigma})} \quad \text{.....(31)}$$

The modified Bessel function of the third kind is given by:

$$K_{\nu}(t)=\frac{1}{2}\int_{0}^{\infty} x^{\nu-1}\exp\left(-\frac{1}{2}t(x+x^{-1})\right)dx \quad \text{.....(32)}$$

Where,

y is the response variable,

μ is the mean response,

σ is the scale parameter and,

ν is the shape parameter.

Zou et al. (2014) analyzed the dispersion term of SI model and compared its performance to the dispersion parameter of NB model in estimating the level of dispersion in the crash data. The dispersion parameter of SI model is given by:

$$h(\sigma,\ \nu)=\frac{2\sigma(\nu+1)}{c}+\frac{1}{c^{2}}-1 \quad \text{.....(33)}$$

It was found that the dispersion term of SI model gives a more reliable estimate of the level of dispersion when compared to the dispersion parameter of NB model (Zou et al., 2014).

The SI and NB models were also compared in terms of hot spot identification using EB estimates by Wu et al. (2013). The SI model performed better than the NB

model when the EB approach was used for hot spot identification. The performance of the SI generalized additive model for location, scale and shape (GAMLSS) for modelling highly dispersed crash was analyzed and was compared to NB and ZINB models. It was found that the SI model performs better when it is used for modelling highly dispersed crash data with long tails.

2.8 NB-Crack Distribution

This distribution was proposed by Saengthong and Bodhisuwan (2012) to model over-dispersed count data. It is a combination of the NB and the Crack (CR) distributions and has a heavy tail. The PMF of NB distribution is the same as before:

$$f(x)=\binom{r+x+1}{x}p^r(1-p)^x \quad \text{.....(34)}$$

Where,

$r > 0$ and $0 < p < 1$.

The PDF of CR distribution is given by (Saengthong and Bodhisuwan, 2012):

$$g(X=x,\lambda,\theta,\gamma)=\frac{1}{\theta\sqrt{2\pi}}\left[\gamma\lambda\left(\frac{\theta}{x}\right)^{\frac{3}{2}}+(1-\gamma)\left(\frac{\theta}{x}\right)^{\frac{1}{2}}\right]\times\exp\left[-\frac{1}{2}\left(\sqrt{\frac{x}{\theta}}-\lambda\sqrt{\frac{\theta}{x}}\right)^2\right],$$

$x>0$ .....(35)

Where,

$\lambda > 0$, $\theta > 0$ and $0 \leq \ \leq 1$.

The definition of Crack distribution as provided by Saengthong and Bodhisuwan (2012) is : Let X be a random variable which follows the negative binomial-Crack distribution with parameters r, λ, θ and γ, X~ NB-CR(r , λ, θ, γ), when X has a NB

distribution with parameter r>0 and p = exp(-a) where a is distributed as CR with positive parameters λ, θ and γ, i.e., X|a~NB(r, p=exp(-a)) and a~CR(λ, θ, γ).

The PMF of NB-CR distribution is given by (Saengthong and Bodhisuwan, 2012):

$$f(X=x,r,\lambda,\theta,\gamma)=\binom{r+x+1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\frac{\exp\left[\lambda\left(1-\sqrt{1+2\theta(r+j)}\right)\right]}{\sqrt{1+2\theta(r+j)}}\times$$

$$\left[1-\gamma\left(1-\sqrt{1+2\theta(r+j)}\right)\right] \quad (36)$$

x = 0, 1, 2, ...

Where,

r, λ, θ > 0 and $0 \leq \gamma \leq 1$.

The NB-CR distribution reduces to negative binomial-inverse Gaussian (NB-IG), negative binomial-Birnbaum-Saunders (NB-BS) and negative binomial-length biased inverse Gaussian (NB-LBIG) for three different cases. Saengthong and Bodhisuwan (2012) also examined the performance of this distribution and compared it to NB and Poisson distributions using real data. It was found that the NB-CR distribution provides a better fit than the NB and Poisson distributions.

2.9 Other Models

This subsection provides a brief description of some other models that have been used by researchers to analyze crash data.

2.9.1 Gamma Count Model

Gamma count model can be used to analyze both over-dispersed and under-dispersed crash data. It was first proposed by Winkelmann (1995). This model was used by Oh et al. (2006) to analyze under-dispersed crash data which consisted of crashes at rail-highway crossings. It was found that the gamma count model could handle such (under-dispersed) data.

The probability for the gamma count model is given by:

$$Pr(y_i = j) = Gamma(\alpha j, \lambda_i) - Gamma(\alpha j+\alpha, \lambda_i) \quad (37)$$

Where,

$\lambda_i = \exp(\beta X_i)$ is the mean of crashes.

$$Gamma(\alpha j, \lambda_i) = 1, \text{ if } j = 0, \quad (38)$$

$$Gamma(\alpha j, \lambda_i) = \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i} u^{\alpha j-1} e^{-u}\, du, \text{ if } j > 0, \quad (39)$$

Where,

α is the dispersion parameter.

If $\alpha > 1$, it implies that over-dispersion exists and if $\alpha < 1$, it implies that under-dispersion exists. α=1 implies that the mean is equal to its variance, in which case the gamma model comes down to the Poisson model.

The CDF (Cumulative Distribution Function) of the model is given by:

$$F(T|\alpha\lambda_i) = \int_0^T \frac{\lambda_i^{\alpha j}}{\Gamma(\alpha j)} u^{\alpha j-1} e^{-\lambda_i u}\, du, \ \alpha>0, \ \lambda_i>0, \ j=0,1,\ldots$$

$$= Gamma(\alpha j, \lambda_i T) \quad (40)$$

The gamma model has certain limitations as one of its assumptions include dependency of observations. According to the gamma model the observation made at time t is dependent on the observation made at time t-1 (Winkelmann, 1995; Cameron, 1998). Since crashes are mostly independent this assumption is not applicable for crash data.

2.9.2 Conway-Maxwell-Poisson Model

The Conway-Maxwell-Poisson (COM-Poisson) distribution was first introduced in 1962 by Conway and Maxwell. It is an extension of the Poisson distribution and is used to handle both under-dispersed and over-dispersed crash data. Shmueli et al. (2005) analyzed different statistical properties of the COM-Poisson distribution. The PDF of COM-Poisson distribution is given by:

$$P(Y=y)= \frac{1}{Z(\lambda, \nu)} \frac{\lambda^y}{(y!)^\nu} \quad \text{.....(41)}$$

$$Z(\lambda,\nu)= \sum_{n=0}^{\infty} \frac{\lambda^n}{(n!)^\nu} \text{ ; } \lambda>0 \text{ and } \nu \geq 0 \quad \text{.....(42)}$$

Where,

Y = crash count,

λ = centering parameter and,

ν = shape parameter.

Over-dispersion is observed when $\nu < 1$, under-dispersion is observed when $\nu > 1$ and the crash data follows Poisson distribution when $\nu = 1$. $\nu = 0$, $\lambda < 1$ would result in geometric distribution and $\nu \rightarrow \infty$ would result in Bernoulli distribution.

The first two moments, mean and variance, were derived by Shmueli et al. (2005) and they are given by:

$$E[Y]= \frac{\partial \log Z}{\partial \log \lambda} \quad \text{.....(43)}$$

$$Var[Y]= \frac{\partial^2 \log Z}{\partial^2 \log \lambda} \quad \text{.....(44)}$$

As the mean and variance for COM-Poisson distribution do not have a closed-form equations, Shamueli et al. (2005) approximated the mean by using an asymptotic expression for Z. The mean and variance thus obtained is given by:

$$E[Y] \approx \lambda^{1/\nu}+\frac{1}{2\nu}-\frac{1}{2} \quad \text{.....(45)}$$

$$Var[Y] \approx \frac{1}{\nu}\lambda^{1/\nu} \quad \text{.....(46)}$$

It was observed that as ν gets closer to 1, λ becomes equal to the mean and as ν becomes small, λ varies greatly from mean. Interpreting the COM-Poisson GLM becomes difficult for over-dispersed data because ν is small for over-dispersed data. Guikema and Coffelt (2008), therefore, introduced an alternative GLM framework to solve this issue. $\lambda^{1/\nu}$ was substituted for μ and the equations are given by:

$$P(Y=y) = \frac{1}{S(\mu,\nu)}\left(\frac{\mu^y}{y!}\right)^y \quad \text{.....(47)}$$

$$S(\mu,\nu) = \sum_{n=0}^{\infty}\left(\frac{\mu^n}{n!}\right)^{\nu} \quad \text{.....(48)}$$

The GLM framework developed by Guikema and Coffelt (2008) based on the above equations can handle both over-dispersed and under-dispersed crash data and it has two links. Its GLM is given by:

$$\ln(\mu) = \beta_{0+} \sum_{i=1}^{p} \beta_i x_i \quad \text{.....(49)}$$

$$\ln(\nu) = \alpha_0 + \sum_{j=1}^{q} \alpha_j z_j \quad (50)$$

Where,

x_i and z_j are the covariates,

p and q are the number of covariates.

Though the parameter estimation for the dual-link COM-Poisson GLM is complex, researchers, i.e. Sellers and Shmueli (2010), have derived its likelihood function which simplified the MLE of the parameters. The performance of COM-Poisson GLM also was examined by Geedipally (2008).

2.10 Negative Binomial-Generalized Exponential Model

The NB-GE distribution was first introduced by Aryuyuen and Bodhisuwan (2013). The mixed distribution combines the NB with the GE distribution. This distribution can handle over-dispersed datasets with a large number of zeroes. The characteristics of this distribution as provided by Aryuyuen and Bodhisuwan (2013) are described below:

The PMF of the NB distribution is:

$$P(Y=y,\mu,\phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)} \left(\frac{\phi}{\mu+\phi}\right)^{\phi} \left(\frac{\mu}{\mu+\phi}\right)^{y} \quad (51)$$

Where,

μ = mean response and,

ϕ = inverse of the dispersion parameter β .

The PDF of generalized exponential distribution is given as follows (Aryuyuen and Bodhisuwan, 2013):

$$f(Z=z,\alpha,\lambda)=\alpha\lambda(1-e^{\lambda z})^{\alpha-1}e^{-\lambda z};\ \alpha,\lambda>0,\ z>0 \quad (52)$$

Where,

α = shape parameter and,

λ = scale parameter.

The exponential distribution is the special case of the GE distribution (i.e., when $\alpha = 1$). The moment generating function of the GE distribution is given as (Aryuyuen and Bodhisuwan, 2013):

$$M_z(t)=(\Gamma(\alpha+1)\Gamma(1-t/\lambda))/\Gamma(\alpha-t/\lambda+1) \quad (53)$$

The mean and variance of the GE distribution are given as (Gupta and Kundu, 1999):

$$E(Z)=\frac{1}{\lambda}(\psi(\alpha+1)-\psi(1)) \quad (54)$$

$$Var(Z)=\frac{1}{\lambda^2}(\psi'(\alpha+1)-\psi'(1)) \quad (55)$$

Where,

$\psi(.)$ = digamma function and,

$\psi'(.)$= derivative of the digamma function $\psi(.)$.

The NB-GE distribution arises by combining the NB and GE distributions, as stated above. The PMF of NB-GE distribution is therefore given as (Aryuyuen and Bodhisuwan, 2013):

$$f(X=x;r,\alpha,\lambda)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\left(\frac{\Gamma(\alpha+1)\Gamma\left(1+\frac{r+j}{\lambda}\right)}{\Gamma\left(\alpha+\frac{r+j}{\lambda}+1\right)}\right);r,\ \alpha,\ \lambda>0 \quad \text{.....(56)}$$

2.11 Estimation Methods

Two methods of estimation, namely, maximum likelihood estimation (MLE) and Bayesian estimation are discussed in this subsection.

Maximum likelihood estimation

This is one of the methods which has been traditionally used to estimate parameters of regression models. According to Casella and Berger (2001), MLE can be defined as follows:

For each sample point y, let $\hat{\beta}(y)$ be a parameter value at which L(β|y) attains its maximum as a function of β , with y held fixed. A maximum likelihood estimator (MLE) of the parameter based on a sample Y is $\hat{\beta}(y)$.

The likelihood function of an independent and identically distributed sample with PDF $f(y|\beta_1,\ldots\beta_k)$ is given by the following equation:

$$L(\beta|y)=L(\beta_1,\ldots,\beta_K|y_1,\ldots..y_n)=\prod_{i=1}^{n} f(y_i|\ \beta_1,\ldots\beta_k) \quad \text{.....(57)}$$

For NB regression model, in order to obtain the coefficients, the first-order should be made equal to zero (Lord and Park, 2013).

The PDF of NB distribution as discussed in section 2.2 is given by:

$$P(Y=y_i,\mu_i,\psi)=\frac{\Gamma(\psi+y_i)}{\Gamma(\psi)\Gamma(y_i+1)}\left(\frac{\psi}{\mu_i+\psi}\right)^{\psi}\left(\frac{\mu_i}{\mu_i+\psi}\right)^{y_i} \quad \text{.....(58)}$$

The two first-order conditions are given by (Lord and Park, 2013):

$$\sum_{i=1}^{n} \frac{y_i-\mu_i}{1+\psi^{-1}\mu_i} x_i=0 \quad (59)$$

$$\sum_{i=1}^{n} \left\{ \frac{1}{(\psi^{-1})^2} \left[\ln\left(1+\psi^{-1}\mu_i\right) - \sum_{j=0}^{y_i-1} \frac{1}{(j+)} \right] + \frac{y_i-\mu_i}{\psi^{-1}(1+\psi^{-1}\mu_i)} \right\} =0 \quad (60)$$

Where,

x_i is a vector of covariates

Bayesian approach

In recent times, researchers have shown more interest in using Bayes approach over MLE to estimate parameters. Statistical software programs, such as WinBUGS and OpenBUGS, use the Bayes approach for parameter estimation. Full Bayes (FB) and EB are two different approaches that have been proposed in highway safety research. The FB approach is more flexible when compared to EB method which makes FB approach easier to use to model crash data (Miranda-Moreno, 2006). Researchers have shown interest in using hierarchical Bayes model to model crash data by Markov Chain Monte Carlo (MCMC) method (Miaou and Song, 2005; Miranda Moreno et al, 2007; Miaou and Lord, 2003). Lord and Park (2013) provided the sampling procedure for MCMC simulation by using slice sampling algorithm within Gibbs sampling. The formulation of the Poisson-Gamma model is given below:

(Likelihood) $y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$

(First Stage) $\lambda_i|\psi \sim \pi_\lambda(\psi)$

(Second Stage) $\psi \sim \pi_\psi(.)$

Where,

$\pi_\lambda(\psi)$ is the prior distribution imposed on the Poisson mean, λ_i,

ψ is a prior parameter, and

$\pi_\psi(\cdot)$ is the hyper-prior on ψ with known hyper-parameters (here, a and b).

The conditional distributions for each parameter for the Poisson-Gamma model are given by (Park, 2010):

$$\pi(\lambda_i|\beta,\psi, y_i)=\text{Gamma}(y_i+\psi, 1+\psi e^{-x_i'\beta}), \text{ for } i=1, 2,\ldots,n \quad (61)$$

$$\pi\left(\beta_j\middle|\lambda,\beta_{-j},\psi\right)=\exp\left\{-\psi\left[\left(\sum_{i=1}^n x_{ij}\right)\beta_j \sum_{i=1}^n \lambda_i e^{-x_i'\beta}\right]\right\}, \text{ for } j=0, 1, \ldots, J \quad (62)$$

$$\pi(\psi|\lambda,\beta, a,b)=\exp\{-n \ln(\Gamma(\psi))+\psi[n \ln(\psi) - \sum_{i=1}^n(x_i'\beta+ \ln(\lambda_i) + \lambda_i e^{-x_i'\beta})]+(a-1) \ln(\psi) -b\psi\} \quad (63)$$

The MCMC sampling procedure using Gibbs sampling, as provided in the Appendix C (Lord and Park, 2013), is as follows:

1. Start with initial values $\lambda^{(0)}$, $\beta^{(0)}$ and $\psi^{(0)}$. Repeat the following steps for t = 1, ..., T_0,....., T_0 +T .

2. Step 1: Conditional on knowing $\beta^{(t-1)}$ and $\psi^{(t-1)}$, draw $\lambda^{(t)}$ from Equation C-29a independently for i =1, 2,...., n .

3. Step 2: Conditional on knowing $\lambda^{(t)}$ and $\psi^{(t-1)}$, draw $\beta^{(t)}$ from Equation C-29b independently for j = 0, 1,...., J using the slice sampling method.

4. Step 3: Conditional on knowing $\lambda^{(t)}$ and $\beta^{(t)}$, draw $\psi^{(t)}$ from Equation C-29c using the slice sampling method.

5. Step 4: Store the values of all parameters (i.e., $\lambda^{(t)}$, $\beta^{(t)}$ and $\psi^{(t)}$). Increase t by one and return to Step 1.

6. Step 5: Discard the first k draws as a burn-in period, where k is defined by the user. The average of the sampled values is calculated for estimating the parameters after equilibrium is reached at the k^{th} iteration.

2.12 Summary

This subsection has documented the results of the literature review on Poisson, NB, ZIP, ZINB, PIG, SI, NB-L, NB-CR, COM-Poisson, gamma count and NB-GE models. Researchers have examined and compared the performance of different models in handling over-dispersed crash data. The above mentioned distributions, their properties and the performance analysis previously documented by other researchers have been summarized in this section. NB-GE is one such distribution which was recently proposed by Bodhisuwan and Aryuyuen (2013) to handle over-dispersed count data. A brief introduction to the distribution and its properties is provided in this section. The next section evaluates the performance of NB-GE distribution and compares it to the Poisson, NB and NB-L distributions using some of the GOF measures

3. PERFORMANCE OF THE NB-GE DISTRIBUTION

This section describes the performance of NB-GE distribution by comparing it to the performance of Poisson, NB and NB-L distributions. In order to compare different distributions, two existing datasets are used. The characteristics of the datasets used are described in the first subsection. GOF measures, such as Chi-squared test and log-likelihood are used to compare the performance of distributions. The second subsection provides a description of GOF measures. R, a statistical software, is used to calculate the parameters of the distribution. The predicted values and GOF statistics for different distributions are then calculated. A brief discussion of the results is presented in the third subsection, and the last subsection summarizes the chapter.

3.1 Description of Datasets

This subsection describes the datasets and their characteristics. The first dataset includes single-vehicle fatal crashes that occurred on divided multilane rural highways between 1997 and 2001. The data was collected as a part of NCHRP 17-29 research project titled "*Methodology for estimating the safety performance of multilane rural highways*" (Lord et al, 2008). The data contained 1,721 segments that varied from 0.10 mile to 11.21 miles, with an average equal to 1.01 miles. The sample mean was equal to 0.13. About 89% of the segments had no fatal crash. The summary statistics of the first dataset is provided in Table 1.

Table 1: Summary statistics of Single-vehicle fatal crashes on divided multilane rural highways between 1997 and 2001 (Lord and Geedipally, 2011)

Crashes	Observed frequency
0	1532
1	162
2	19
3	6
4+	2

Table 2: Summary statistics of Single-vehicle roadway departure crashes on rural two-lane horizontal curves between 2003 and 2008 (Lord and Geedipally, 2011).

Crashes	Observed Frequency
0	29087
1	2952
2	464
3	108
4	40
5	9
6	5
7	2
8	3
9	1
10+	1

The second dataset includes single-vehicle roadway departure fatal crashes that occurred on 32,672 rural two-lane horizontal curves between 2003 and 2008. The sample mean is equal to 0.14. For this dataset, about 90% of the data experienced no crashes during the 5-year period. The summary statistics for the second dataset is provided in Table 2.

3.2 Goodness of Fit

This subsection describes the definitions of GOF measures used to compare distributions. Two performance measures were used to compare the GOF of different distributions: Pearson's Chi-Squared statistic and Log-likelihood value. The Log-likelihood is calculated as the logarithm of likelihood for each observation. The Chi-squared and log-likelihood values are calculated using the following expressions.

$$\text{Chi-squared} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{.....(64)}$$

$$\text{Log-likelihood} = \sum_{i=1}^{n} \text{Log}(P_i) \quad \text{.....(65)}$$

Where,

O_i is the observed frequency for i^{th} observation,

E_i is the expected frequency for i^{th} observation,

P_i is the expected likelihood for i^{th} observation, and

N is the total number of observations.

The smaller the log-likelihood and Chi-squared value, the better the fit of the distribution for the data.

3.3 Results and Discussion

This subsection describes the comparison analysis among the Poisson, NB, NB-L and NB-GE distributions. Statistical software, R is used to find the parameter values for NB-GE distribution. The code used in R is provided in Appendix A which was developed by Aryuyuen and Bodhisuwan (2013). The parameter values and the GOF measures for Poisson, NB, NB-GE ad NB-L distributions is taken from the research paper by Geedipally et al., (2011). The GOF analysis, based on the Chi-Square and log-likelihood, are presented in Tables 3 and 4.

From Table 3 it can be observed that NB-GE distribution predicts frequency closer to the observed frequency when compared to other models. The Chi-square and log-likelihood values obtained also suggest that NB-GE distribution provides a better fit for the crash data.

Table 3. Single-Vehicle Fatal Crashes on Divided Multilane Rural Highways Between 1997 and 2001

Crashes	**Observed Frequency**	**Poisson**	**NB**	**NB-GE**	**NB-L**
0	1532	1509.2	1534.4	1532.6	1532.9
1	162	198	154.7	158.9	158.3
2	19	13.0	25.8	23.6	23.7
3	6	0.6	4.9	4.5	4.6
4	2	0	1.2	1.8	1.4
Parameters		μ=0.131	μ=0.131 ϕ=0.434	r=1.28 α=1.5 β=13.569	Θ=1532.9 R=1.851
Chi-square		102.99	2.73	**1.39**	1.68
Log-likelihood		-715.1	-696.1	**-694.5**	-695.6

Note: Bold characters indicate that the distribution is a better fit

Figure 1 provides a comparison of the zero counts predicted using different distributions.

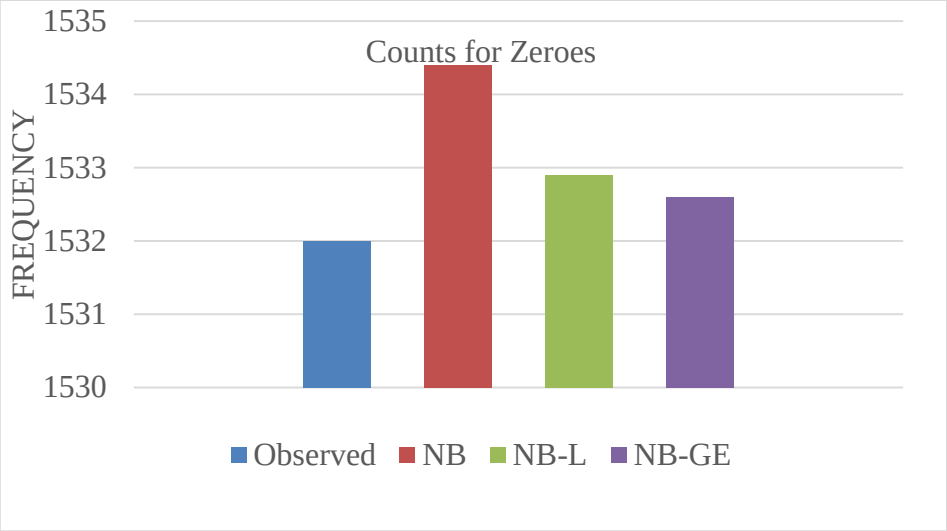

Fig 1. Predicted and observed zero counts for the first dataset

It can be observed that NB overestimates the zeroes in the data set and NB-GE distribution predicts zero counts closer to the observed values when compared to NB-L and NB distributions. Table 4 provides a comparison of the estimates and GOF statistics for different distributions for the second dataset.

Table 4. Single-Vehicle Roadway Departure Crashes on Rural Two-Lane Horizontal Curves between 2003 and 2008

Crashes	Observed Frequency	Poisson	NB	NB-GE	NB-L
0	29087	28471.6	29204.8	29097.8	29133.6
1	2952	3918.0	2706	2908.4	2855.5
2	464	269.6	567	498.3	503.1
3	108	12.4	141.1	115.9	120.9
4	40	0.4	37.8	34.3	35.9
5	9	0.0	10.6	11	13.1
6	5	0.0	3.0	4.1	3.3
7	2	0.0	0.9	3	3.3
8	3	0.0	0.3	0.9	0.0
9	1	0.0	0.1	0.4	3.0
10	1	0.0	0.0	0.2	3.3
Parameters		μ=0.138	μ=0.138 ϕ=0.284	r=0.937 α=1.280 β=8.999	Θ=9.212 R=1.018
Chi-square		2297.31	57.47	**6.38**	11.68
Log-likelihood		-14,208.1	-13,557.7	**-13,525**	-13,529.8

Note: Bold characters indicate that the distribution is a better fit

From the results obtained in Table 4 it can be observed that the NB-GE distribution predicts the crash frequency closer to the observed frequency when compared to the other distributions for this crash data set. The GOF statistics calculated also suggest that NB-GE distribution provides a better fit. Figure 2 provides a comparison of the zero count estimates for different distributions.

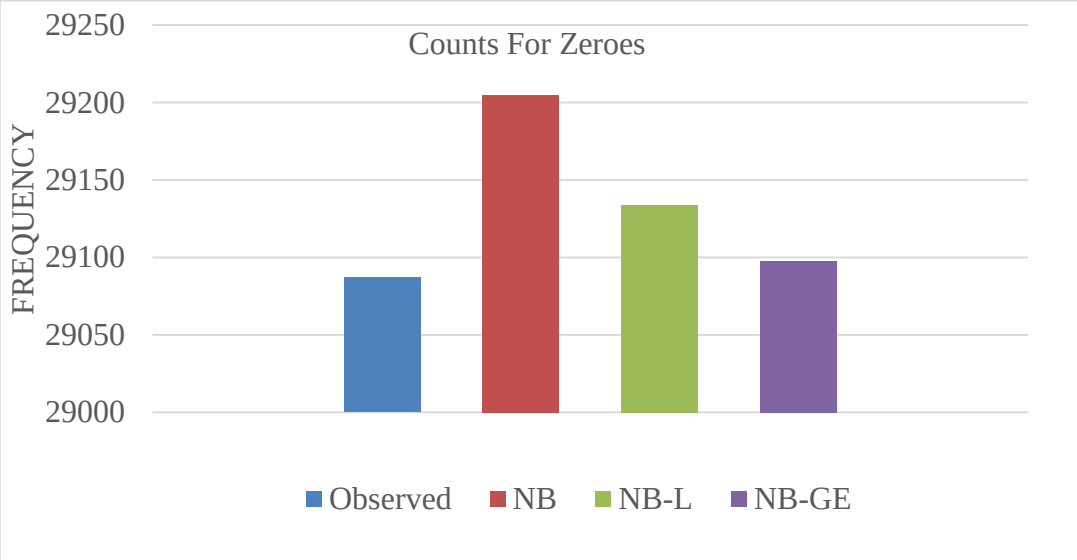

Fig 2. Predicted and observed zero counts for the second dataset

From the provided above for the second crash dataset, it can be observed than NB overestimates zero counts. NB-GE predicts the counts closer to the observed values when compared to other distributions.

3.4 Summary

This subsection provides a brief summary of this section.

In most cases, crash data exhibit over-dispersion. In order to predict crashes, a model which can handle over-dispersion should be used. The NB distribution can handle over-dispersed data whereas the Poisson distribution can handle data sets whose variance is equal to mean. But when the over-dispersed crash data contains excessive zero counts, the NB distribution tends to overestimate the number of zero counts. This section examined the performance of NB-GE distribution in handling over-dispersed crash datasets containing excess zeroes and compared its performance to Poisson, NB and NB-GE distributions.

In order to calculate the predicted crashes for NB-GE distribution, its parameters had to be estimated. The parameters of NB-GE distribution were estimated using the statistical computing software, R. The Chi-square test and the log-likelihood value were calculated in order to test the GOF of the distribution. These values when compared to the results obtained for Poisson, NB and NB-L distributions suggested that both NB-GE and NB-L perform much better than Poisson and NB distributions. Also, when NB-GE distribution was compared to NB-L it was observed that NB-GE distribution performs slightly better than NB-L distribution. It was also observed that NB distribution over-estimated the zero counts for both datasets. The next section documents the application of NB-GE GLM for two observed crash datasets and compares its performance to other models.

4. APPLICATION OF THE NB-GE GENERALIZED LINEAR MODEL FOR OVER-DISPERSED CRASH DATA

Regression models help in establishing relationship between the roadway characteristics and crashes. Different models have been proposed to handle over-dispersed data with large number of zeroes. Some of the models that have already been used to analyze such data include ZIP and NB, SI, PIG and NB-L models. For this thesis, the performance of NB-GE model will be compared with the NB and NB-L models. The first part of this section discusses the development of NB-GE GLM. The second part of this section gives a summary of the datasets that have been used to analyze the performance of NB-GE model. The third part of this section gives a brief description of various GOF measures that are used in this research to compare the performance of NB-GE model followed by results and summary of the section.

4.1 NB-GE Generalized Linear Model

The NB-GE distribution, as the name suggests, is a combination of NB and GE distributions.

As an alternative parameterization, the NB-GE distribution can be written as:

$$P(X = x, \mu, \varphi, \alpha) = \int NB(x; \varphi, z\mu\mu)GE(z\alpha, \lambda)\ dz \quad \text{.....(66)}$$

Here, μ is a function of the covariates and z has a Generalized Exponential distribution.

For the GLM of NB-GE distribution, the mean response of crashes is obtained by multiplying E(z) with μ and μ is considered to have a log linear relationship with the covariates.

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{q} \beta_i X \quad \text{.....(67)}$$

Where,

X= traffic and geometric variables

β_s = regression coefficients to be estimated

q = total number of covariates in the model.

The mean response of the number of crashes at a particular site is given by:

$$E(X) = \mu \times E(Z) = e^{\sum_{i=1}^{q} \beta_i X + \beta_0} \times \frac{1}{\lambda}(\psi(\alpha+1) - \psi(1)) \quad \text{.....(68)}$$

The coefficients α, λ and ϕ, of the above model will be estimated in OpenBUGS. OpenBUGS is preferred over WinBUGS as the GE distribution is readily available in OpenBUGS. In order to get the estimates of the coefficients, three Markov chains with 50,000 iterations are used. The first 40,000 iterations are then discarded and the last 10,000 iterations are considered for the analysis.

4.2 Description of Datasets

The first dataset used for this purpose was collected in Indiana over a five-year period at 338 road sections of a rural interstate and the second dataset was collected in Michigan over one-year period (2006) at 33,970 road sections of a two-lane rural highway. These datasets are used by Geedipally et al. (2012). The characteristics of these data sets are given in Tables 5 and 6.

Table 5. Summary Statistics for the Indiana Data (Geedipally et al., 2012)

Variable	Min.	Max.	Average (std. dev)	Total
Number of Crashes (5 years)	0	329	16.97 (36.30)	5737
Average daily traffic over the 5 years (ADT)	9442	143,422	30237.6 (28776.4)	--
Minimum friction reading in the road segment over the 5-year period (FRICTION)	15.9	48.2	30.51 (6.67)	--
Pavement surface type (1 if asphalt, 0 if concrete) (PAVEMENT)	0	1	0.77 (0.42)	--
Median width (in feet) (MW)	16	194.7	66.98 (34.17)	--
Presence of median barrier (1 if present, 0 if absent) (BARRIER)	0	1	0.16 (0.37)	--
Interior rumble strips (RUMBLE)	0	1	0.72 (0.45)	--
Segment length (in miles) (L)	0.009	11.53	0.89 (1.48)	300.09

Table 6. Summary Statistics for the Michigan Data (1996) (Geedipally et al., 2012)

Variable	Min.	Max.	Average (std. dev)	Total
Number of Crashes (1 year)	0	61	0.68 (1.77)	23168
Annual average daily traffic (AADT)	160	20,994	4507.5 (3280.6)	--
Segment length (L) (miles)	0.001	54.54	0.18 (0.58)	6212
Shoulder width (in feet) (SW)	0	24	16.94 (5.26)	--
Lane width (in feet) (LW)	8	15	11.22 (0.78)	--
Speed limit (SPEED) (mph)	25	55	52.47 (6.39)	--

4.3 Goodness of Fit

Three performance measures were used to test the GOF of the GLM of NB-GE. A brief description of those three criteria is given below:

CURE plot

The cumulative residual (CURE) plot is used to examine the fitting of the model with respect to each covariate (Hauer and Bamfo, 1997). The closer the curve is to zero the better a model fits the data.

Mean Absolute Deviation (MAD)

It is calculated by taking the average of the absolute deviations. The closer its value is to zero the better the model. It's given by the following equation (Oh et al., 2003):

$$MAD=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad \text{.....(69)}$$

Mean Squared Predictive Error (MSPE)

It is calculated by taking the average of the square of the absolute deviations. The model that has MSPE value closer to zero is a better model when compared to the other models. It is given by the following equation (Oh et al., 2003):

$$MSPE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|^2 \quad \text{.....(70)}$$

4.4 Results and Discussion

This subsection presents preliminary results obtained for the NB-GE GLM. Both Indiana data and Michigan data were analyzed in the OpenBUGS software to obtain the coefficients for the NB-GE GLM. Criteria such as MAD, MSPE, Deviance Information Criteria (DIC) and CURE plots were used to assess the models. The results obtained are summarized in the tables and plots given below. Table 7 provides the modelling results and some GOF statistics for Indiana data.

From the values for GOF measures shown in Table 7, it can be seen that NB-L and NB-GE perform better that NB model. The DIC, MAD and MSPE values are higher for NB model when compared to the other two. When NB-L and NB-GE is compared it can be observed that NB-L seems to perform slightly better than NB-GE.

Cure plots for the variables ADT and friction are provided in figures 3 and 4 for Indiana data.

Table 7. Modeling Results for the Indiana Data.

Variable	NB		NB-L		NB-GE	
	Value	**Std. dev**	**Value**	**Std. dev**	**Value**	**Std. dev**
INTERCEPT (β_0)	-4.779	0.979	-3.739	1.115	-3.233	0.423
Ln(ADT) (β_1)	0.7219	0.091	0.630	0.106	0.570	0.067
FRICTION (β_2)	-0.02774	0.008	-0.02746	0.011	-0.0284	0.010
PAVEMENT (β_3)	0.4613	0.135	0.4327	0.217	0.481	0.158
MW (β_4)	-0.00497	0.001	-0.00616	0.002	-0.00651	0.002
BARRIER (β_5)	-3.195	0.234	-3.238	0.326	-3.240	0.324
RUMBLE (β_6)	-0.4047	0.131	-0.3976	0.213	-0.349	0.154
α[1]	0.934	0.118	0.238	0.083	2.339	0.427
λ					1.526	0.284
DIC	1,900		**1,701**		1,784	
MAD[2]	6.91		**6.89**		7.04	
MSPE[3]	206.76		**195.54**		202.93	

Note: Bold characters indicate a better fit

From the table and CURE plot of the Indiana data for ADT variable, it can be observed that NB-L and NB-GE models perform much better than NB model. For the CURE plots, the residuals are adjusted such that the final values are zero. Also, it was

observed that both NB-L and NB-GE predict total crashes at all sites better than NB model.

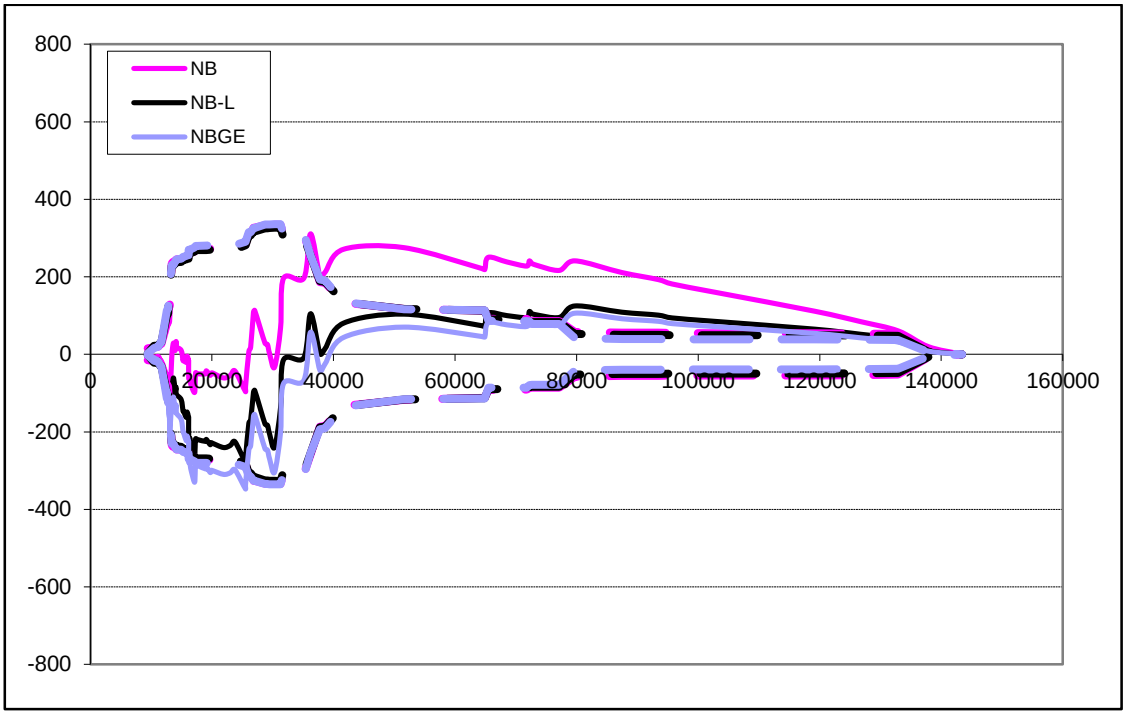

Fig 3. Cumulative Residual Plot for Indiana Data (ADT variable)

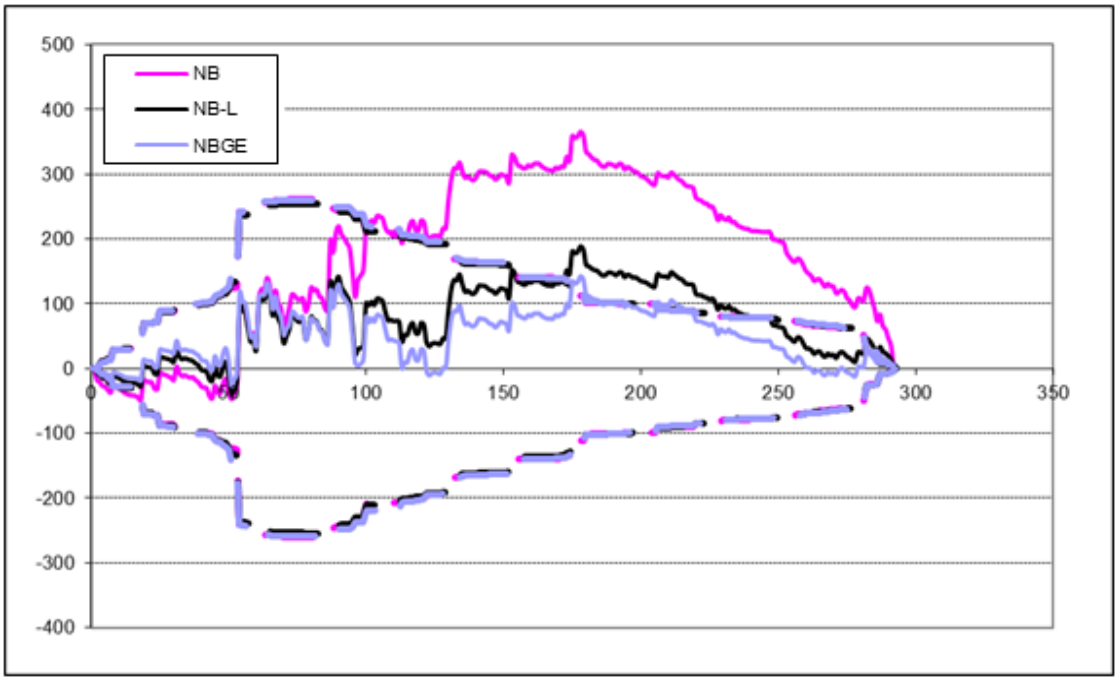

Fig 4. Cumulative Residual Plot for Indiana Data (Friction variable)

It can be observed in Figure 4 that the CURE plot of Indiana data for friction variable also provides similar results. NB-GE and NB-L seem to perform better than NB model. The NB-GE also shows better performance when compared to NB-L when CURE plot for friction variable is considered.

Table 8 below provides the modeling and GOF results for the Michigan data.

Table 8. Modeling Results for the Michigan Data.

Variable	NB		NB-L		NB-GE	
	Value	**Std. dev**	**Value**	**Std. dev**	**Value**	**Std. dev**
INTERCEPT (β_0)	-3.412	0.239	-3.2607	0.193	-1.58	0.1724
Ln(AADT) (β_1)	0.4267	0.014	0.4243	0.015	0.4212	0.011948
L (β_2)	0.9571	0.009	0.9615	0.009	0.9705	0.0087
SW (β_3)	-0.00009	0.002	-0.0003	0.002	0.0007	0.0023
LW (β_4)	0.0589	0.013	0.0508	0.011	0.0374	0.0105
SPEED (β_5)	0.0098	0.002	0.0091	0.002	0.0071	0.0019
α^1	0.5727	0.019	0.1024	0.002	2.829	0.1462
λ					7.18	1.457
DIC	59,354		**56,046**		57,670	
MAD[2]	0.651		**0.648**		0.6498	
MSPE[3]	2.831		2.884		3.089	

Note: Bold characters indicate a better fit

From the GOF values in Table 8, it can be observed that MAD is lower for both NB-L and NB-GE models when compared to NB model. The NB-L model seems to perform slightly better than the NB-GE model. The CURE plots for two variables ADT and length segment are given below.

Figures 5 and 6 provide CURE plots of Michigan data for AADT variable and the length segment variable respectively. From the CURE plot of Michigan data for AADT variable, it can be observed that NB-GE model performs better than NB and NB-L models as the curve is closer to the axis. The second CURE plot is plotted for the length segment variable. It can be observed that NB-GE and NB-L seem to have lower residuals when compared to NB model.

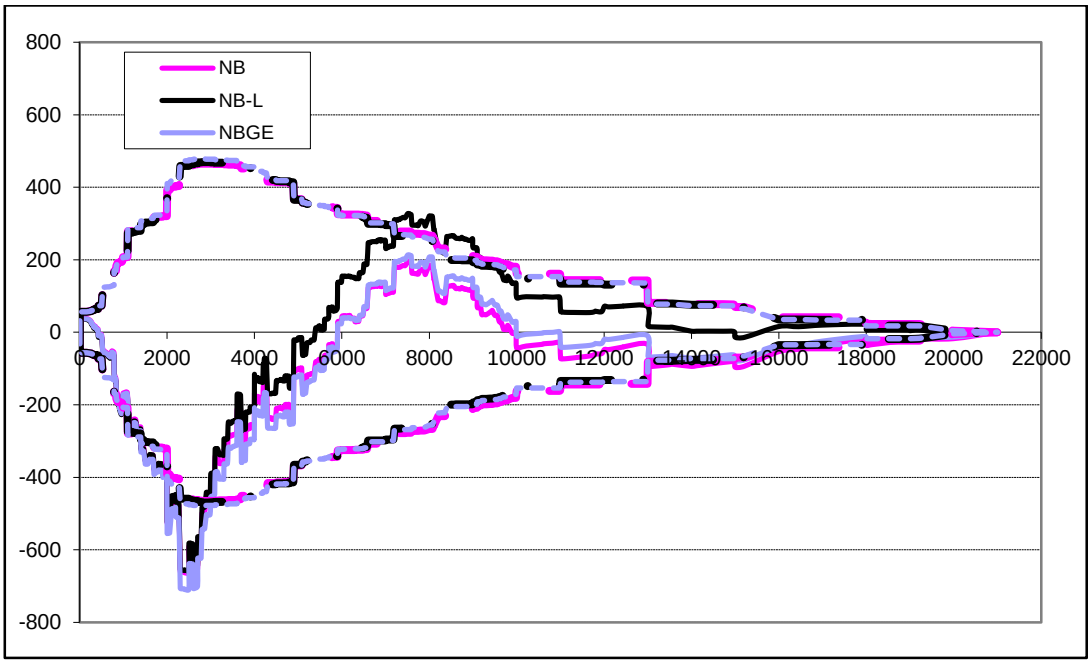

Figure 5. Cumulative Residual Plot for Michigan Data (AADT variable)

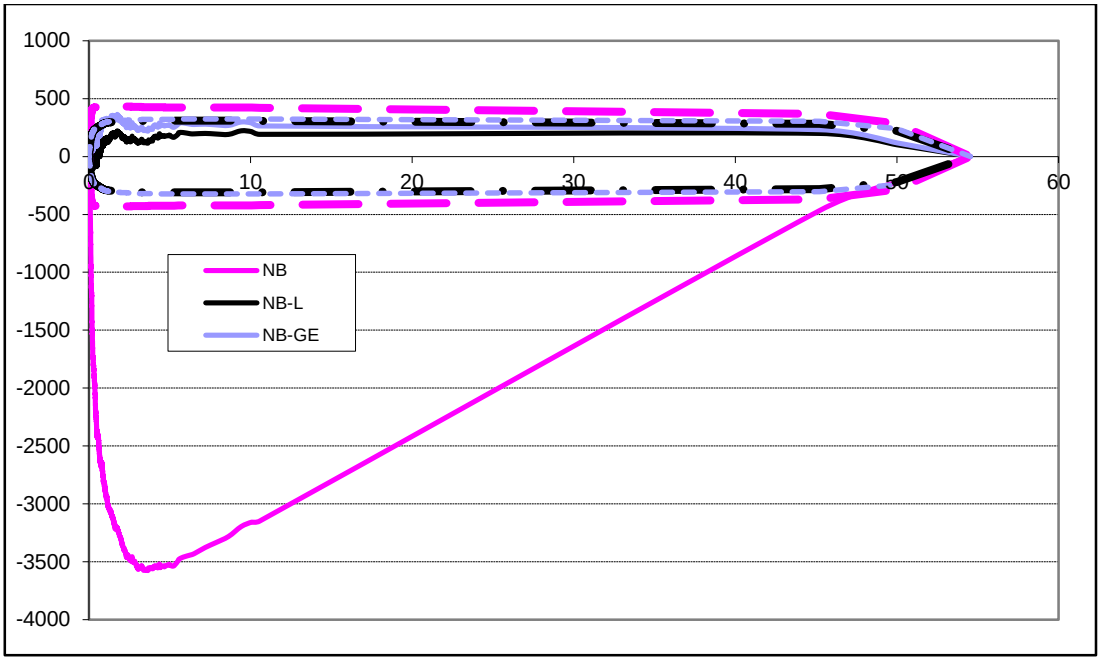

Figure 6. Cumulative Residual Plot for Michigan Data (length segment)

The parameterization of the NB-GE used in this work is slightly different than the parameterization described by Aryuyuen and Bodhisuwan (2013), but is similar to the one proposed for the NB-L in Geedipally et al. (2011). When a GLM is considered with the original formulation proposed by Aryuyuen and Bodhisuwan (2013), the mean response is a non-linear, non-invertible function of the covariates and the parameters, which makes it difficult to characterize the predicted response. On the contrary, the parameterization proposed in this paper is easily interpretable.

Despite the nice interpretability offered by this characterization, MCMC chains could still suffer from poor mixing. This often results from the fact that the $GE(\alpha, \lambda)$ distribution behaves quite differently for α <=1 and α >1. Moreover, the mean of the GE tends to infinity as α increases for a given fixed value of λ. This problem can be

mitigated by restricting the parameter α over some range. In this paper, a prior of the form $\alpha \sim$ uniform (1, 3) was used. In addition, when α and λ of the GE distribution were varied and α was plotted against the density f(x), the following observations were made: 1) for $\alpha < 0.5$, f(x) is concentrated around 0; 2) for $\alpha > 1$, f(x) has different support depending on the scale parameter, is unimodal. Thus, the range (1, 3) for α and (1, 2) for λ was considered reasonable.

Since the GE distribution currently exists in the OpenBUGS software, the NB-GE model estimation can be easily implemented. However, the computational time for MCMC runs was slightly longer than the NB model because it involves an additional parameter compared to the NB model. But, the difference in computational times between the two models was not very large. The code used in OpenBUGS are provided in Appendix A for reference.

4.5 Summary

This section discussed the development of the GLM for NB-GE distribution and application of the GLM for analyzing crash data characterized by an excess of zeros. The NB-GE was compared to the NB and the recently introduced three-parameter model, NB-L, using a total of four datasets. The analyses was carried using two different datasets and the GLM was estimated using OpenBUGS.

From the Table 7 and CURE plot for the Indiana data, it was observed that the NB-L and NB-GE models perform much better than NB model. For the CURE plots, the

residuals were adjusted such that the final values are zero. Also, it was observed that both NB-L and NB-GE predict total crashes at all sites better than NB model.

From the CURE plot for Michigan data it seems that NB model has small residuals than NB-L and NB-GE models. However, it was observed that NB-L and NB-GE models predict total crashes at all sites better than NB model.

The results obtained using Indiana and Michigan data sets show that both NB-GE and NB-L models perform much better than NB model. The NB-GE was also easier to implement than the NB-L model, which may make this model more useful.

The next section covers the performance of NB-GE GLM using simulated data. Datasets with different levels of dispersion and percentages of zeroes are simulated to analyze the performance of GLM of the NB-GE distribution. The performance of NB-GE model is compared to NB model for hot spot identification.

5. PERFORMANCE OF NB-GE GENERALIZED LINEAR MODEL FOR SIMULATED CRASH DATA

This section covers the performance analysis of both NB and NB-GE models. The next task for this research is to examine the characteristics of NB-GE model for different percentages of zeroes in the data and determine the threshold that would favor the NB-GE model over the NB model. To accomplish this objective, the ranking of sites for the hotspot identification will be used as a performance measure. Data needs to be simulated in order to perform this task.

The first subsection deals with simulation protocol. Second subsection of this section describes the performance measures used for comparative analysis followed by discussion of the results obtained and summary of the section.

5.1 Data Simulation

Data with varying percentage of zeroes and different dispersion values is simulated using the following steps:

1. Using simulated covariates and assumed parameter values, the "true" means for a given sample size is calculated.
2. All the sites are ranked based on their true mean
3. Later, using an assumed dispersion parameter, the crash counts are simulated with varying percentages of zeroes i.e., 50%, 60%, 70%, 80%, 90%. Two levels of dispersions (i.e. low and high) were considered. Thus, 10 different situations are evaluated.

4. The NB and NB-GE models are then be fitted and the parameters re-estimated. Steps 3 and 4 will be repeated 10 times.

5. The average of the means obtained from the simulation are then calculated. The sites are then ranked and the difference in ranks between the predicted and the observed will be examined. The sites which have a difference in ranks greater than 30 will be considered to be misidentified. NB and NB-GE models are thus compared using performance measures, such as the false discovery rate, percentage of false negatives etc., among others.

5.2 Performance Measures

In order to compare the performance of different models, different performance measures have been used. The performance measures that are used for this research have already been discussed by Park et al. (2014) and Wu et al. (2013).

There are several possible outcomes when a site is classified based on a hot spot identification method as provided in the table below (Miranda-Moreno, 2006).

Table 9. Possible outcomes of classification (Miranda-Moreno, 2006)

	Number of sites "detected" as non-hotspots	Number of sites "detected" as hotspots	
Number of "true" non-hotspot	U	V	n_0
Number of "true" hotspot	R	S	n_1
	n-D	D	N

Where,

n is the total number of sites in the set under analysis,

n_0 is the number of "true" non-hotspots ,

n_1 is the number of "true" hotspots ,

U is the number of sites correctly classified as non-hotspots,

V is the number of false positives or Type I errors,

R is the number of false negatives or Type II errors,

S is the number of sites correctly classified as hotspots, and

D is the number of sites detected hotspots as hotspots.

False Discovery Rate (FDR): The ratio of false positives among all the detected hotspots by a model. Smaller FDR value indicates a better model.

FDR=V/D.....(71)

False Negative Rate (FNR): The ratio of false negatives among all the detected non-hotspots by a model. Smaller FNR value indicates a better model

$FNR=R/(n-D)$.....(72)

Sensitivity (SENS): The ratio of correctly detected hotspots by a model among the true hotspots. Larger SENS value indicates a better model.

$SENS=s/n_1$.....(73)

Specificity (SPEC): The ratio of correctly detected non-hotspots by a model among the true non-hotspots. Larger SPEC value indicates a better model

$SPEC=U/n_0$.....(74)

Risk (RISK): The ratio of total number of false positives and false negatives among all the sites under analysis. Smaller value indicates a better model.

$RISK=(V+R)/n$.....(75)

The Percentage of false negative (P_{FN}) and percentage of false positive (P_{FP}) are given by the following equations

$$P_{FN}=\frac{N_{FN}}{N_{TS}} X\ 100 \quad \text{.....(76)}$$

$$P_{FP}=\frac{N_{FP}}{N_{TH}} X100 \quad \text{.....(77)}$$

Where,

N_{FN} is the number of false negatives,

N_{FP} is the number of false positives,

N_{TS} is the number of truly safe sites, and

N_{TH} is the number of truly hazardous sites.

5.3 Results and Discussion

The coefficients of the NB and NB-GE models were estimated using OpenBUGS. An extension of R called R2OpenBUGS was used in which OpenBUGS could be called through R. The data was simulated using R, then OpenBUGS was called out through R to estimate coefficients for the simulated data. In order to get the estimates of the coefficients, three Markov chains with 50,000 iterations were used. The first 40,000 iterations were discarded and the last 10,000 iterations were considered for the analysis. The summary statistics of the simulated data are provided in Appendix B. Table 10 provides a summary of mis-specified sites for 10 different scenarios of simulated data

Table 10. Number of mis-specified sites for different scenarios

Scenario	NB	NB-GE
50% zero counts; Low dispersion	2	2
50% zero counts; High dispersion	17	13
60% zero counts; Low dispersion	0	3
60% zero counts; High dispersion	8	12
70% zero counts; Low dispersion	0	1
70% zero counts; High dispersion	16	19
80% zero counts; Low dispersion	59	59
80% zero counts; High dispersion	30	24
90% zero counts; Low dispersion	186	180
90% zero counts; High dispersion	84	73

It can be observed from Table 10 that NB-GE performs better than NB model when the percentage of zeroes in the dataset is greater than 80% for both low and high dispersion. Hence, NB-GE model might provide a better fit in terms of hot spot identification when the crash data has zero counts higher than 80%. The tables below provide performance measures for different scenarios of simulated data. Three threshold values; 90^{th}, 85^{th} and 80^{th} percentiles were considered to evaluate the performance measures

Table 11. Performance measures for 50% zero counts and low dispersion

	NB			**NB-GE**		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.020	0.042	0.031	0.020	0.042	0.031
FNR	0.002	0.007	0.007	0.002	0.007	0.007
SENS	0.980	0.960	0.970	0.980	0.960	0.970
SPEC	0.998	0.994	0.992	0.998	0.994	0.992
RISK	0.004	0.012	0.012	0.004	0.012	0.012
P_{FN}	0.22	0.706	0.75	0.22	0.706	0.75
P_{FP}	2	4	3	2	4	3

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 12. Performance measures for 50% zero counts and high dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.136	0.087	0.075	0.136	0.087	0.075
FNR	0.013	0.014	0.018	0.013	0.014	0.018
SENS	0.880	0.92	0.93	0.880	0.92	0.93
SPEC	0.888	0.986	0.982	0.888	0.986	0.982
RISK	0.024	0.024	0.028	0.024	0.024	0.028
P_{FN}	1.333	1.412	1.75	1.333	1.412	1.75
P_{FP}	12	8	7	12	8	7

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

From Tables 11 and 12 above it can be observed that for 50% zero count both NB and NB-GE perform similarly at both low and high dispersion.

Tables 13 and 14 provide performance measures when the percentage of zeroes in the dataset is 60. NB and NB-GE models seem to perform equally well except for 2 cases. At low dispersion NB-GE shows a better performance when the threshold is 80^{th} percentile and at high dispersion NB-GE model performs better when the threshold is 90^{th} percentile.

Table 13. Performance measures for 60% zero counts and low dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.020	0.087	0.020	0.020	0.087	**0.010**
FNR	0.002	0.014	0.005	0.002	0.014	**0.002**
SENS	0.980	0.92	0.98	0.980	0.92	**0.990**
SPEC	0.998	0.986	0.995	0.998	0.986	**0.997**
RISK	0.004	0.024	0.008	0.004	0.024	**0.004**
P_{FN}	0.22	1.412	0.5	0.22	1.412	**0.25**
P_{FP}	2	8	2	2	8	**1**

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 14. Performance measures for 60% zero counts and high dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.064	0.042	0.053	**0.021**	0.042	0.064
FNR	0.067	0.007	0.013	**0.004**	0.007	0.015
SENS	0.940	0.960	0.950	**0.960**	0.960	0.940
SPEC	0.993	0.994	0.987	**0.995**	0.994	0.985
RISK	0.012	0.012	0.020	**0.008**	0.012	0.024
P_{FN}	0.667	0.706	1.25	**0.444**	0.706	1.5
P_{FP}	6	4	10	8	4	6

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 15. Performance measures for 70% zero counts and low dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.021	0.042	0.031	**0.020**	0.042	0.042
FNR	0.004	0.007	0.007	**0.002**	0.007	0.010
SENS	0.960	0.960	0.970	**0.980**	0.960	0.920
SPEC	0.995	0.994	0.992	**0.998**	0.994	0.990
RISK	0.008	0.012	0.012	**0.004**	0.012	0.016
P_{FN}	0.444	0.706	0.75	**0.22**	0.706	1
P_{FP}	8	4	3	**2**	4	8

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 16. Performance measures for 70% zero counts and high dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.020	0.056	0.042	0.020	0.056	0.042
FNR	0.002	0.009	0.010	0.002	0.009	0.010
SENS	0.980	0.947	0.920	0.980	0.947	0.920
SPEC	0.998	0.990	0.990	0.998	0.990	0.990
RISK	0.004	0.016	0.016	0.004	0.016	0.016
P_{FN}	0.22	0.941	1	0.22	0.941	1
P_{FP}	2	5.333	8	2	5.333	8

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

From tables 15 and 16, it can be observed that both NB and NB-GE seem to perform equally well at 70% zeroes except in one case. For low dispersion when the threshold is 90^{th} percentile, NB-GE seems to perform better than NB model.

Table 17. Performance measures for 80% zero counts and low dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.163	0.103	0.1	0.163	0.103	**0.099**
FNR	0.016	0.017	0.025	0.016	0.017	**0.023**
SENS	0.860	0.907	0.900	0.860	0.907	**0.910**
SPEC	0.984	0.983	0.975	0.984	0.983	**0.977**
RISK	0.028	0.028	0.04	0.028	0.028	**0.036**
P_{FN}	1.556	1.647	2.500	1.556	1.647	**2.250**
P_{FP}	14	9.333	10	14	9.333	**9**

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

From tables 17 and 18 it can be observed that when the simulated crash data contains 80% zero counts, the NB-GE seems to perform slightly better than NB model. For low dispersion, both perform equally well when the threshold is 90^{th} and 85^{th} percentile. But when the threshold is 80^{th} percentile NB-GE model seems to perform better. At high dispersion NB-GE performs better than NB model at both 85^{th} and 80^{th} percentile threshold values.

Table 18. Performance measures for 80% zero counts and high dispersion

	NB			NBGE		
Percentile	90^{th}	85^{th}	80^{th}	90^{th}	85^{th}	80^{th}
FDR	0.021	0.087	0.053	0.021	**0.071**	**0.042**
FNR	0.004	0.014	0.013	0.004	**0.012**	**0.010**
SENS	0.960	0.900	0.900	0.960	**0.933**	**0.920**
SPEC	0.995	0.975	0.987	0.995	**0.988**	**0.990**
RISK	0.008	0.024	0.020	0.008	**0.02**	**0.016**
P_{FN}	0.444	1.412	1.25	0.444	**1.176**	**1**
P_{FP}	8	8	10	8	**6.667**	**8**

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 19. Performance measures for 90% zero counts and low dispersion

	NB			NBGE		
Percentile	90th	85th	80th	90th	85th	80^{th}
FDR	0.25	0.136	0.136	**0.219**	0.136	0.136
FNR	0.023	0.022	0.031	**0.020**	0.022	0.031
SENS	0.8	0.88	0.88	**0.82**	0.88	0.88
SPEC	0.978	0.979	0.970	**0.980**	0.979	0.970
RISK	0.04	0.036	0.048	**0.036**	0.036	0.048
P_{FN}	2.222	2.118	3	**2**	2.118	3
P_{FP}	20	12	12	**18**	12	12

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

Table 20. Performance measures for 90% zero counts and high dispersion

	NB			NBGE		
Percentile	90th	85th	80th	90th	85^{th}	80^{th}
FDR	0.111	0.107	0.075	**0.064**	**0.087**	**0.064**
FNR	0.111	0.019	0.018	**0.067**	**0.014**	**0.015**
SENS	0.9	0.893	0.930	**0.940**	**0.92**	**0.940**
SPEC	0.989	0.981	0.982	**0.993**	**0.986**	**0.985**
RISK	0.02	0.032	0.028	**0.012**	**0.024**	**0.024**
P_{FN}	1.111	1.882	1.75	**0.667**	**1.412**	**1.5**
P_{FP}	10	10.667	7	**6**	**8**	**6**

Note: Smaller FDR, FNR, RISK, P_{FN}, P_{FP} are better and larger SENS and SPEC values are better

From tables 19 and 20, for 90% zero count in simulated crash data, NB-GE performs better than NB model for all threshold values when the dispersion is high. At low dispersion NB-GE model performs better than NB model when the threshold is 80^{th} percentile. Therefore, NB-GE seems to be a better choice over NB model when the percentage of zeroes in the crash data is greater than 80.

5.4 Summary

This subsection provides a brief summary of this chapter and results obtained in the previous section.

In order to examine the performance of NB and NB-GE models for crash data with different characteristics, data was simulated with varying zero percentages and dispersion levels. The parameters were then estimated to get the predicted values. Statistical softwares, R with an extension R2OpenBUGS and OpenBUGS were used to achieve this.

Both NB and NB-GE seem to perform the same way when percentage of zero counts in the crash data is low. In this analysis, the performance measures considered were ranking sites for hot-spot identification, FDR, FNR, SENS, SPEC, RISK, P_{FN}, P_{FP}. The number of mis-specified sites for NB and NB-GE are almost equal when the percentage of zeroes in the crash data is less than 80%. For percentages higher than 80 (here 90%) it was observed that NB-GE model performs better than the NB model for both low and high dispersion.

Therefore, it can be argued that NB-GE has an advantage over NB model when the number of zero counts in the crash data is higher than 80%. For lower percentages, NB model would be a wiser choice as it is easier to implement and consumes less time when compared to NB-GE model.

6. SUMMARY AND RECOMMENDATIONS

This section provides a brief summary of the research work done for this thesis and also describes few recommendations for future work.

6.1 Summary

The NB-GE distribution was recently introduced by two researchers, Aryuyuen and Bodhisuwan (2013). It is a 3-parameter distribution and is a mixture of both NB and GE distributions.

This thesis has described the development and application of the NB-GE distribution and its GLM for analyzing crash data characterized by a high dispersion with an excess of zeros.

The first part of the thesis covered the analysis of the performance of NB-GE distribution by comparing its performance to the Poisson, NB and NB-L distributions using two over-dispersed crash data sets containing excess zeroes. The Pearson's Chi-squared test and log-likelihood value were used as performance measures. It was found that both NB-L and NB-GE distributions fit the data better than the NB and Poisson distribution and NB-GE seems to perform slightly better than NB-L.

The next part of the thesis documented the development of the GLM for NB-GE distribution and examined its performance using two datasets based on data collected in Indiana and Michigan by comparing it to NB and NB-L models. MAD, MSPE and CURE plots were used as the criteria to evaluate the performance. It was found that in the case of the Indiana data, the tables and the CURE plot showed that that NB-GE and

NB-L models provided better statistical performance than the NB model. The CURE plots showed that the total crashes predicted by NB-GE and NB-L models are closer to the observed values when compared to NB model.

The last part of the analysis consisted in finding the point where NB-GE model starts performing better than NB model using simulated data. Ranking sites for hot-spot identification was used as the performance measure. Data were simulated for varying zero percentages and dispersion. R and OpenBUGS were used to simulate data and estimate parameters. It was found that when the percentage of zeroes in the data set exceeds 80 NB-GE model ranks the sites better than NB model. Therefore, it would be recommended to use the NB model when the data set contains lower percentage ($\leq 80\%$) zeroes as it is easier and simpler to use. But if the percentage of zeroes exceeds 80%, the NB-GE model would be a better choice.

6.2 Recommendations

This subsection provides some recommendations for further work

- It was observed that both NB-L and NB-GE models seem to perform almost equally well, at least for the Indiana and Michigan data. Further work can be done to determine the performance of these models in ranking sites for hot-spot identification based on the characteristics of the dataset. Their performances can be compared similar to that done in this thesis for NB and NB-GE models.
- Crash data are usually characterized by small sample size and low sample means. Further work can be done to determine the effect of these characteristics on the

NB-GE model. Work on developing and optimizing the MLE for the NB-GE should also be examined so that the NB-GE could be more easily used by transportation safety analysts and other researchers.

- A GLM for NB-CR distribution can also be developed and its performance can be compared to the NB-GE and NB-L models.

REFERENCES

Abdel-Aty, M., Addella, M.F. (2004). Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data. Transportation Research Record, Issue 1897, pp. 106–115.

Aryuyuen. S., Bodhisuwan, W. (2013). Negative binomial-generalized exponential (NB-GE) distribution. Applied Mathematical Sciences, Vol. 7, No. 22, pp. 1093-1105

Cameron, A.C., Trivedi, P.K., 1998. Regression analysis of count data. Cambridge University Press, Cambridge, UK

Casella, G., Berger, R., 2001. Statistical inference, second ed. Duxbury, Pacific Grove, CA.

Clark, S.J., Perry, J.N., 1989. Estimation of the negative binomial parameter k by maximum quasi-likelihood. Biometrics 45, pp. 309-316.

Conway, R.W., Maxwell, W.L. (1962). A queuing model with state dependent service rates. Journal of Industrial Engineering, Vol. 12, pp. 132-136

Efron, B. (1986). Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association, Vol. 81, No. 395, pp. 709-721.

Geedipally, S.R., 2008. Examining the application of Conway-Maxwell-Poisson model for analyzing traffic crash data. Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University, College Station, Texas.

Geedipally, S.R., Lord, D., Dhavala, S.S. (2012). The Negative binomial-Lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention, Vol. 45, No. 2, pp. 258-265.

Golob, T., and Recker, W. (1987). An analysis of truck-involved freeway accidents using log-linear modeling. Journal of Safety Research Vol. 18, No. 3, pp. 121–136.

Guikema, S.D., Coffelt, J.P., 2008. A flexible count data regression model for risk analysis. Risk Analysis Vol.28, No.1, pp. 213-223

Gupta, R.D., Kundu, D. (1999). Generalized exponential distributions, Australia & New Zealand Journal of Statistics, Vol. 41, No. 2, pp. 173–188.

Hallmark, S.L., Qiu, Y., Pawlovitch, M., McDonald, T.J. (2013). Assessing the safety impacts of paved shoulders. Journal of Transportation Safety and Security, Vol. 5, No. 2, pp. 131-147.

Hauer, E., 1997. Observation before-after studies in road safety: Estimating the effect of highway and traffic engineering measures on road safety. Elsevier Science Ltd., Oxford.

Hilbe, J.M. (2007). Negative binomial regression. Cambridge University Press, Boston, MA.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 3 (4), pp. 53-57.

Lambert, D., 1992. Zero-inflated poisson regression, With an application to defects in manufacturing. Technometrics 34 (1), pp. 1-14

Lord, D., Geedipally, S.R., 2011. The Negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis and Prevention Vol.43, No. 5, pp. 1738-1742.

Lord, D., Geedipally, S.R., Guikema, S., 2010. Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion. Risk Analysis, Vol. 30, No. 8, pp. 1268-1276.

Lord, D., Guikema, S.D., Geedipally, S. (2008a). Application of the Conway-Maxwell Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention, Vol. 40, No. 3, pp. 1123–1134.

Lord, D ., Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, Vol. 44, No. 5, pp. 291-305

Lord, D., Park, P.Y-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis and Prevention Vol.40, No. 4, pp. 1441-1457.

Lord, D., Washington, S.P., Ivan, J.N. (2005). Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accident Analysis & Prevention, Vol. 37, No. 1, pp. 35-46.

Lord, D., Washington, S.P., Ivan, J.N. (2007). Further notes on the application of zero inflated models in highway safety. Accident Analysis & Prevention, Vol. 39, No. 1, pp. 53-57.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, Vol. 10, pp. 325-337.

Maher, M. (1991). A new bivariate negative binomial model for accident frequencies. Traffic Engineering and Control, Vol. 32, pp. 422–425.

Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention, Vol. 26, No. 4, pp. 471–482.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Research Record 1840, pp. 31-40.

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial

dependence. Accident Analysis and Prevention, Vol. 37, No. 4, pp. 699–720.

Miranda-Moreno, L.F., 2006. Statistical models and methods for the identification of hazardous locations for safety improvements. Ph.D. thesis, Department of Civil Engineering, University of Waterloo, Canada.

Miranda-Moreno, L., Lord, D., Fu, L., 2007. Evaluation of alternative hyper-priors for Bayesian road safety analysis. Paper 08-1788. In: Proceedings of the 84th Annual Meeting of the Transportation Research Board, Washington, D.C.

Mitra, S., Chin, H.C., Quddus, M.A., 2002. Study of intersection accident by maneuvre type. Transportation Research Record 1784, pp. 43-50.

Oh, J., Washington, S.P., Nam, D. (2006). Accident prediction model for railway-highway interfaces. Accident Analysis and Prevention, Vol. 38, No. 2, pp. 346-56.

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J. (2003). Validation of the FHWA crash models for rural intersections: Lessons Learned. Transportation Research Record 1840, pp. 41-49.

Park, E.S., Lord, D. (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the

Transportation Research Board (2019), TRB, National Research Council Washington, DC, pp. 1–6.

Qin, X., Ivan, J.N., Ravi shanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention Vol. 36, No. 2, pp. 183-191.

Saha, K., Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics, Vol.61, No. 3, pp. 179-185.

Shankar, V., Mannering, F., and Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention, Vol. 27, No. 3, pp. 371–389.

Shankar, V., Milton, J., Mannering, F. (1997). Modeling accident frequency as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention, Vol. 29, No. 6, pp. 829-837.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergal, M.B. (2003). Modeling crashes involving pedestrians and motorized traffic. Safety Science, Vol. 41, No. 7, pp. 627-640.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society, Part C, Vol. 54, pp. 127-142.

Venables, W.N., Smith, D.M., The R development team, 2005. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Insightful Corporation, Seattle, WA.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16, pp. 275-289.

Winkelmann, R., 1995. Duration dependence and dispersion in count-data models. Journal of Business Economic Statistics V0l. 13, No. 4, pp. 467-474.

Zha, L., Lord., D., Zou, Y.(2014). The Poisson Inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data (to be published in Journal of Transportation Safety and Security).

Zou, Y., Geedipally, S.R., Lord, D. (2013). Evaluating the double Poisson generalized linear model. Accident Analysis & Prevention, Vol. 59, pp. 497-505.

Zou, Y., Lord, D., Zhang., Y.(2011). Analyzing highly dispersed crash data using the Sichel generalized additive models for location, scale and shape (working paper).

Zou, Y., Wu, L., Lord, D. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models (to be published in AMAR).

APPENDIX A

Code for MLE of NB-GE distribution

NB-GE code used in R for MLE of NB-GE distribution is provided below for reference. It was taken from the research paper by Aryuyuen and Bodhisuwan (2013).

```
mlogl<-function(theta,x){
fnbge<-function(theta,x){
mm<-length(x)
 k<-numeric(mm)
nbge<-function(theta,x){
 if(x==0){
p<-(-log(factorial(theta[1]+x-1))+log(factorial(theta[1]-1)) -log(gamma(theta[2]+1))-
log(gamma(1+(theta[1]/theta[3]))) +log(gamma(theta[2]+(theta[1]/theta[3])+1)))}
else if(x>0){
pp1<-
(gamma(theta[2]+1)*(gamma(1+theta[1]/theta[3])))/(gamma(theta[2]+theta[1]/theta[3]+
1))
 for(j in 1:x){
p1<-((factorial(x)/(factorial(j)*factorial(x-j)))*(-1)^j)
*(gamma(theta[2]+1))*((gamma(1+(theta[1]+j)/theta[3]))
/(gamma(theta[2]+(theta[1]+j)/theta[3]+1))) pp1<-pp1+p1}
p<-(-log(factorial(theta[1]+x-1)))+log(factorial(theta[1]-1)) +log(factorial(x))-log(pp1)}
p}
```

```
for(i in 1:length(x)){
k[i]<-nbge(theta,x[i])}k} sum(fnbge(theta,x))}
theta.start<-c(1,1,1)
out<-nlm(mlogl, theta.start, x=x)
r_MLE<-out$estimate[1]
a_MLE<-out$estimate[2]
b_MLE<-out$estimate[3]
```

Codes used for NB ad NB-GE GLM

The NB model code used in OpenBUGS for Michigan and Indiana datasets are provided below.

NB model code:

```
model
{for(i in 1:500)
  { NB_rand[i] ~ dnegbin(p[i],r)
 p[i] <- r/(r+mu[i])
      log(mu[i]) <-
b0+b1*S.lnF[i]+b2*S.x6[i]+b3*S.x28[i]+b4*S.x31[i]+b5*S.x32[i]+b6*S.x38[i]}
  b0 ~ dnorm(-5, 0.1)
    b1 ~ dnorm(0, 0.1)
    b2 ~ dnorm(0, 0.1)
    b3 ~ dnorm(0, 0.1)
    b4 ~ dnorm(0, 0.1)
```

b5 ~ dnorm(0, 0.1)

b6 ~ dnorm(0, 0.1)

r~ dgamma(0.01, 0.01)}

NB-GE model code:

```
model
{for(i in 1:100)
  {
    NB_rand[i]~ dnegbin(p[i],r)
  p[i] <- r/(r+a[i]*mu[i])
        log(mu[i]) <-
b0+b1*S.lnF[i]+b2*S.x6[i]+b3*S.x28[i]+b4*S.x31[i]+b5*S.x32[i]+b6*S.x38[i]
    a[i] ~ dgen.exp(alpha,beta)}
 b0~ dnorm(0, 0.01)
 b1~ dnorm(0, 0.1)
 b2~ dnorm(0, 0.1)
 b3~ dnorm(0, 0.1)
 b4~ dnorm(0, 0.1)
 b5~ dnorm(0, 0.1)
 b6~ dnorm(0, 0.1)
        r~ dgamma(2,1)
       alpha~ dunif(1, 3)
       beta~dunif(1, 2)}
```

APPENDIX B

Modelling results of simulated data

The average values of the parameter estimates for simulated data are given below

Table 21. Modelling results for simulated data with 50% zeroes

	Low Dispersion		High Dispersion	
Parameters	NB	NB-GE	NB	NB-GE
β_0	-5.41161	-4.99603	-5.35453	-4.56416
β_1	0.631672	0.592625	0.703972	0.64514
β_2	-0.02372	-0.02593	-003191	-0.03385
β_3	0.433452	0.40842	0.447161	0.420685
β_4	-0.00575	-0.00603	-0.00567	-0.00574
β_5	-3.16862	-3.36031	-3.17565	-3.19817
β_6	-0.36308	-0.41185	-0.27942	-0.30404
alpha		2.790784		2.417716
beta		1.558889		1.544692

Table 22. Modelling results for simulated data with 60% zeroes

	Low Dispersion		High Dispersion	
Parameters	NB	NB-GE	NB	NB-GE
β_0	-5.97349	-5.77643	-5.07352	-4.23458
β_1	0.665512	0.636317	0.634425	0.572577
β_2	-0.02865	-0.03123	-0.02869	-0.03044
β_3	0.415015	0.420257	0.455837	0.428905
β_4	-0.00467	-0.00499	-0.00422	-0.00436
β_5	-3.21902	-3.41927	-2.90849	-2.84642
β_6	-0.41751	-0.46318	-0.35321	-0.37324
alpha		2.726861		1.775373
beta		1.533271		1.530838

Table 23. Modelling results for simulated data with 70% zeroes

	Low Dispersion		High Dispersion	
Parameters	NB	NB-GE	NB	NB-GE
β_0	-6.31988	-5.56489	-5.58843	-4.73716
β_1	0.654702	0.585182	0.634512	0.565181
β_2	-0.02642	-0.02963	-0.02509	-0.02641
β_3	0.423776	0.439319	0.525508	0.466496
β_4	-0.00549	-0.00566	-0.00505	-0.00443
β_5	-2.75323	-3.15508	-3.14307	-3.1103
β_6	-0.35401	-0.39289	-0.42405	-0.46993
alpha		2.459398		1.871667
beta		1.527845		1.516769

Table 24. Modelling results for simulated data with 80% zeroes

	Low Dispersion		High Dispersion	
Parameters	NB	NB-GE	NB	NB-GE
β_0	-6.61041	-6.96244	-5.67929	-5.11561
β_1	0.634091	0.670127	0.61055	0.567376
β_2	-0.03302	-0.03354	-0.02933	-0.03145
β_3	0.398229	0.415342	0.460032	0.47168
β_4	-0.00634	-0.0064	-0.00487	-0.00424
β_5	-3.19	-3.42269	-3.40217	-3.60223
β_6	-0.23027	-0.26561	-0.45388	-0.50905
alpha		2.154115		1.919416
beta		1.520861		1.527086

Table 25. Modelling results for simulated data with 90% zeroes

	Low Dispersion		High Dispersion	
Parameters	NB	NB-GE	NB	NB-GE
β_0	-5.12029	-5.92836	-5.81137	-5.10108
β_1	0.457991	0.536374	0.605935	0.551628
β_2	-0.0452	-0.04341	-0.03633	-0.03812
β_3	0.612016	0.545095	0.410079	0.312055
β_4	-0.00836	-0.00789	-0.0083	-0.00814
β_5	-2.51513	-2.72246	-3.19961	-3.14555
β_6	-0.43503	-0.44047	-0.39312	-0.37582
alpha		2.006068		1.908525
beta		1.500885		1.510854